

# A genome-wide screen reveals new regulators of the 2-cell-like cell state

**Pierre-Antoine Defossez** (✉ [pierre-antoine.defossez@u-paris.fr](mailto:pierre-antoine.defossez@u-paris.fr))

Université Paris-Cité, CNRS, Epigenetics and Cell Fate Unit

**Nikhil Gupta**

Université Paris-Cité, CNRS, Epigenetics and Cell Fate Unit

**Lounis Yakhou**

Université Paris-Cité, CNRS <https://orcid.org/0000-0003-4907-3780>

**Julien Richard Albert**

Université Paris-Cité, CNRS, Institut Jacques Monod <https://orcid.org/0000-0001-9950-9461>

**Fumihito Miura**

Department of Biochemistry, Kyushu University Graduate School of Medical Sciences

<https://orcid.org/0000-0003-2656-486X>

**Laure Ferry**

Université Paris-Cité, CNRS, Epigenetics and Cell Fate Unit

**Olivier Kirsh**

<https://orcid.org/0000-0001-6200-5681>

**Anaëlle Azogui**

Université Paris-Cité, CNRS, Epigenetics and Cell Fate Unit

**Kohsuke Yamaguchi**

Université Paris-Cité, CNRS, Epigenetics and Cell Fate Unit

**Marthe Laisné**

Université Paris-Cité, CNRS, Epigenetics and Cell Fate Unit

**Cécilia Domrane**

Université Paris-Cité, CNRS, Epigenetics and Cell Fate Unit

**Frédéric Bonhomme**

Institut Pasteur

**Arpita Sarkar**

Université Côte d'Azur, Inserm, CNRS, IRCAN

**Marine Delagrangue**

Institut de Biologie de l'École Normale Supérieure (IBENS)

**Bertrand Ducos**

Institut de Biologie de l'École Normale Supérieure (IBENS)

**Gael Cristofari**

Institute for Research on Cancer and Aging of Nice (IRCAN) <https://orcid.org/0000-0001-5620-3091>

**Takashi Ito**

Department of Biochemistry, Kyushu University Graduate School of Medical Sciences

<https://orcid.org/0000-0001-6097-2803>

**Maxim Greenberg**

Institut Jacques Monod <https://orcid.org/0000-0001-9935-8763>

---

**Article****Keywords:**

**Posted Date:** June 3rd, 2022

**DOI:** <https://doi.org/10.21203/rs.3.rs-1561018/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published at Nature Structural & Molecular Biology on July 24th, 2023. See the published version at <https://doi.org/10.1038/s41594-023-01038-z>.

# Abstract

In mammals, only the zygote and blastomeres of the early embryo are fully totipotent. This totipotency is mirrored *in vitro* by mouse "2-cell-like cells" (2CLCs), which appear at low frequency in cultures of Embryonic Stem cells (ESCs). Because totipotency is incompletely understood, we carried out a genome-wide CRISPR KO screen in mouse ESCs, searching for mutants that reactivate the expression of *Dazl*, a robust 2-cell-like marker. Using secondary screens, we identify four mutants that reactivate not just *Dazl*, but also a broader 2-cell-like signature: the E3 ubiquitin ligase adaptor SPOP, the Zinc Finger transcription factor ZBTB14, MCM3AP, a component of the RNA processing complex TREX-2, and the lysine demethylase KDM5C. Functional experiments show how these factors link to known players of the 2 cell-like state. These results extend our knowledge of totipotency, a key phase of organismal life.

## Introduction

Embryonic development begins with the zygote: a single, totipotent cell which can give rise to all cells of the embryo. In mouse, this totipotency is still present in the 2-cell embryo, as each cell can regenerate a whole embryo and its extra-embryonic tissues, but this capacity quickly wanes with ensuing divisions<sup>1,2</sup>. Coinciding with the totipotency of the 2-cell blastomeres is the process of Zygotic Gene Activation (ZGA), i. e. the initiation of transcription in the embryo. A key driver of the ZGA is the transcriptional regulator DUX (*Dux* in mouse/*DUX4* in human), which binds to MERVL repeated elements and activates their transcription<sup>3-5</sup>. This, in turn, promotes a general opening of chromatin<sup>4</sup>, as well as the induction of the 2-cell embryo transcriptional signature, containing genes such as *Zscan4a-f*<sup>6</sup>. By the 4-cell stage, the totipotency and the high chromatin accessibility regress, via mechanisms that are only partially understood. The study of totipotency has suffered from technical limitations as early embryos can only be obtained in limited number, but the situation has greatly improved with the discovery of the following *in vitro* system.

Mouse embryonic stem cells (ESCs)—like their counterparts in the Inner Cell Mass of the blastocyst—are pluripotent but not totipotent. However, totipotent-like cells do spontaneously arise in ESC culture in serum, where they can constitute ~ 0.1–0.5% of the population<sup>7</sup>. These cells, termed "2-cell-like cells" (2CLCs) due to their shared properties with 2-cell blastomeres, are now well recognized as a tractable system allowing to get insight into the biology of the 2-cell stage embryo<sup>8,9</sup>.

2CLCs have specific chromatin features and transcriptional regulation. Compared to ESCs, 2CLCs have ~ 30% less DNA methylation, more active chromatin marks, and increased histone mobility. DUX orchestrates the 2CLC transcriptome, as it does in the embryo<sup>3-5</sup>. Other key actors are the related transcriptional activators DPPA2 and DPPA4<sup>10,11</sup>, as well as the tumor suppressor p53<sup>12</sup>, which all act upstream of DUX. Transcriptional hallmarks of the 2CLC state include the derepression of MERVL repeats, and activation of multiple genes also expressed in the 2-cell embryo, including *Zscan4a-f*<sup>6</sup> and *Zfp352*. Some genes robustly expressed in 2-cell embryos and in 2CLCs become re-expressed in later

developmental contexts, such as gametogenesis<sup>13</sup>. This is the case, for instance, of *Asz1*, *Spz1*, and of *Dazl*, the latter of which encodes an RNA-binding protein essential for spermatogenesis<sup>14</sup>. This partial overlap between 2-cell-like and germline transcriptomes comes in part from the existence of shared repressive mechanisms that are lifted in both situations, namely DNA methylation and histone modification by the non-canonical Polycomb Repressive Complex 1.6 (ncPRC1.6)<sup>15–18</sup>. A consequence of this situation is that regulators of *Dazl* could potentially be relevant for germline development, in the 2-cell embryo, or both.

Given the biological importance of totipotency, its relevance for basic research, and its potential interest for medicine, efforts have been made by the community to understand the emergence of 2CLCs. Candidate analyses have shown that chromatin opening is a key factor in the reprogramming of ESCs to 2CLCs<sup>19</sup>. In addition, some medium to high-throughput screens have been undertaken, providing important contributions to our understanding of this biological system<sup>15,20–22</sup>. However, these screens have been limited in several important ways: they have focused on the reactivation of the MERVL repeats while neglecting other markers of totipotency. In addition, several of these screens used targeted siRNA screening, which does not cover the whole genome. Finally, some of these studies have employed gain-of-function approaches which may induce biases.

To extend our knowledge of the 2-cell-like state, we have undertaken an original, genome-wide, loss-of-function genetic screen in mouse ESCs. We used the reactivation of *Dazl* as a 2CLC marker and a dual-positive selection scheme based on antibiotic resistance and FACS. This led to the identification of 40 high-confidence hits, including many of the already known actors in the DNA methylation and ncPRC1.6 pathways. Secondary screens based on the expression of additional 2CLC markers allowed us to identify the hits that induce a 2CLC state, of which we characterized four —SPOP, ZBTB14, MCM3AP/GANP, KDM5C— in more detail by transcriptomics and epigenomics. Our functional experiments then establish the epistasis of these factors relative to known regulators. Among our findings, we show that the lysine demethylase KDM5C shares most of its targets with ncPRC1.6 and represses the 2CLC signature in a catalytically-independent manner. Together our data bring new concepts and new actors to our comprehension of the 2-cell-like state.

## Results

### Design and validation of the epigenetic reporter

We first re-analyzed published expression data<sup>19,23</sup>, which confirmed that *Dazl* is highly expressed in 2-cell-like-cells, but not in serum-grown ESCs (Fig. 1A). Switching ESCs from serum to 2i conditions, which is known to dramatically reduce DNA methylation<sup>24,25</sup>, also results in *Dazl* re-expression (Figure S1A). Mining expression data collected through mouse development, we observed that the abundance of *Dazl* mRNA in 2i-grown ESCs is comparable to that seen in late 2-cell embryos, whereas the abundance in serum-grown ESCs is roughly equivalent to that seen in 8-cell-embryos (Fig. 1B).

Two chromatin-based mechanisms are already known to repress *Dazl* transcription in ESCs grown in serum: DNA methylation and recruitment of ncPRC1.6; inactivating either pathway by removing some of the essential actors (Fig. 1C) causes the reactivation of *Dazl* expression in serum-grown cells<sup>16–18,26,27</sup>. To identify additional factors repressing *Dazl* gene expression, directly or indirectly, we reasoned that we could carry out a CRISPR KO screen on serum-grown ESCs, selecting for cells in which *Dazl* becomes reactivated (Fig. 1C). Such factors could be gene-specific, acting only on *Dazl*, or they could be reactivating wider transcriptional programs, of which *Dazl* expression is only one component. Specifically, these programs could mirror the 2CLC state, gametogenesis, or both. In addition, some KOs might act by causing the cells to become 2i-like. Secondary screens would be required to distinguish these possibilities (Fig. 1C).

Before embarking on the screen, a number of controls were performed. Using WGBS, RNA-seq, and RT-qPCR, we verified that in our ESC background, J1, *Dazl* was expressed and its promoter unmethylated in 2i conditions, whereas *Dazl* was repressed, its promoter methylated in serum condition (Figure S1A-B).

As *Dazl* fulfilled all the conditions to be a useful marker of the 2-cell-like state, we then used CRISPR/Cas9 to knock in 2 selectable markers into the gene: mScarlet, a bright red fluorescent protein, and Hygro<sup>R</sup>, which confers resistance to Hygromycin (Figs. 1D, S1C). The two markers were inserted in-frame within the *Dazl* coding sequence, at exon 6 (which is ~ 6 kb away from the promoter and present in all splicing forms), and the two reporter proteins were separated by T2A and P2A self-cleaving peptides (Figs. 1D, S1C). After recombinant clones were picked and characterized by genomic PCR and sequencing, one of the correct integrants was further analyzed, and droplet digital PCR (ddPCR) was used to establish that only one mScarlet-Hygro<sup>R</sup> insertion had occurred in the genome (Figure S1D). To summarize, we generated a heterozygous J1-based mouse ESC line in which one allele of *Dazl* is wild-type, and the other allele contains the mScarlet-Hygro<sup>R</sup> insertion, which inactivates that allele (Fig. 1F). We will refer to this line as the DASH (*Dazl*-Scarlet-Hygro) reporter line.

As expected, DASH cells in 2i are mScarlet-positive (~ 75% over threshold) and Hygromycin-resistant (to a dose > 125 µg/ml), but they are mScarlet-negative and Hygromycin-sensitive in serum condition (Figs. 1E-F, S1E). Also as expected, MeDIP confirmed that the *Dazl* promoter was methylated in serum and unmethylated in 2i (Figure S1F). All prerequisites having been met, we then used the DASH line for a genome-wide CRISPR KO screen.

## Genome-wide KO screening yields a list of 40 high-confidence hits

We grew the DASH cells in serum and infected them with a lentiviral library of ~ 80,000 vectors co-expressing Cas9 and sgRNAs targeting ~ 20,000 genes<sup>28</sup>. The coverage was ~ 150 infected cells per sgRNA, and two independent screens were carried out in parallel. After an initial Puromycin selection to eliminate non-infected cells, Hygromycin selection was applied in two steps of increasing concentration (Fig. 2A). At the threshold we used, mScarlet-positive cells represented ~ 3% of the starting DASH

population; this proportion increased to ~ 25–30% in the Hygromycin-resistant cells obtained after selection (Figure S2A), and the Hygromycin resistant/mScarlet-positive cells were purified by FACS. In this sorted population, RT-qPCR revealed a number of differences relative to the starting population: upregulation of *Dazl* (as expected), but also increased expression of *Prdm14*, and decreased expression of *Dnmt3a*, *Dnmt3b*, and *Dnmt3l*, all of which are characteristic of 2i-like cells<sup>24,25</sup> (Figure S2B). The bulk DNA methylation in this selected population was lower than in the sample recovered before selection (Figure S2C), and the methylation of the *Dazl* promoter (assessed by MeDIP) was also lower than in the pre-selection sample, indicating that some cells in the populations have lost DNA methylation locally and/or globally (Figure S2D).

The sgRNAs present in the Hygro-resistant/mScarlet-positive cells were amplified and sequenced, and the data analyzed statistically using MAGeCK<sup>29</sup>. This procedure yielded an ordered list of candidates, ranked by *p*-value (Fig. 2B). The top 40 candidates had a *p*-value smaller than  $5.10^{-4}$  and were analyzed further (Fig. 2B). GO terms enriched in this set of candidates include “Maintenance of DNA methylation”, along with “Glycosaminoglycan synthesis” and “Heparan sulfate synthesis” (two related terms), and “FGFR signaling pathway” (Fig. 2C). As for Uniprot GO terms, “Repressor”, “Chromatin regulator”, and “DNA binding” were enriched (Figure S2E).

We then grouped the 40 candidates based on their known functions and their interactions in the STRING database (Fig. 2D). We recovered 3 factors required for DNA methylation maintenance<sup>30</sup> (UHRF1 ranked #2, DNMT1 ranked #11, USP7 ranked #28), as well as components of ncPRC1.6 (E2F6 ranked #4, MGA ranked #5, TFDP1 ranked #38); these hits were expected (Fig. 1C) and validate the screening procedure. In addition, we obtained 8 candidates in the “TGFβ-Wnt signaling” and the “FGFR signaling” clusters, genes involved in “RNA processing” —including three out of four subunits of the TREX-2 complex (ENY2 ranked #24, SEM1 ranked #30, MCM3AP ranked #34)—, a lysine demethylase, KDM5C (ranked #16), two components of E3-ubiquitin ligase complexes (SPOP<sup>31</sup> ranked #8, FBXW7 ranked #25) as well as additional cytoplasmic (PDCL ranked #22, DPYSL4 ranked #35) and nuclear factors (ZBTB14, ranked #18).

New infections were used to generate CRISPR KO populations for 22 selected genes in the top 40, of which 20 increased the number of mScarlet-positive cells in the population (Fig. 2E). This 91% rate of true positives establishes the robustness of the screen results.

## Refining the list to potential 2CLC regulators

We then wanted to narrow the focus of our search. One class of hits we expected to recover from the screen, but that would not be relevant to the 2CLC state, were KOs causing the cells to be blocked in differentiation, causing them to be “2i-like” even in serum condition, and therefore to have low DNA methylation and high *Dazl* expression. In support of this possibility, mutation of genes in the “Heparan Sulfate/Glycosaminoglycan synthesis” pathway decreases FGF/MAPK signaling, and thus increase the proportion of naïve cells<sup>32</sup>; therefore the hits in this cluster likely affect the differentiation process and

were not considered for further analysis here. For the same reason, the "FGFR" and "TGF $\beta$ -Wnt signaling" clusters were also discarded. To identify direct regulators of gene expression, we chose to focus on nuclear factors, leaving aside proteins acting in the cytoplasm or at the ER membrane. We did not pursue further proteins connected to DNA methylation or ncPRC1.6, which have already been extensively studied<sup>15,20</sup>. On the basis of DNA methylation and gene expression analyses, we selected 4 of the remaining candidates for further experiments: the lysine demethylase KDM5C; the E3 ubiquitin ligase adaptor SPOP; the Zinc Finger and BTB-containing protein ZBTB14; and MCM3AP (the homolog of human GANP<sup>33</sup>), the scaffolding subunit of TREX-2, a complex that couples gene expression and mRNA export<sup>34-36</sup>. To the best of our knowledge, none of these factors have previously been shown to regulate the 2-cell-like-state.

We carried out one additional control to rule out the possibility that the selected KOs are in a pseudo-2i state, in which they would fail to respond to serum and differentiation cues. For this, we let the serum-grown cells undergo spontaneous differentiation by removing LIF. While *Tcf711* mutant cells, used as a control, failed to differentiate, in all cases, the colonies of these 4 candidates (*Mcm3ap*, *Spop*, *Kdm5c*, and *Zbtb14*) were rounded in serum and upon LIF removal became irregularly shaped, a mark of differentiation (Figure S2F). In addition, during spontaneous differentiation, the KO cells of these 4 candidates repressed pluripotency genes (*Prdm14*, *Pou5f1*) and induced the differentiation marker *Fgf5* to a similar extent as their WT counterpart (Figure S2G). These data rule out the possibility that the KO cells are locked in a 2i-like state. We therefore continued our characterization of the 4 candidates, *Mcm3ap*, *Spop*, *Kdm5c*, and *Zbtb14*.

## Generation, characterization, and rescue of individual KO clones

To better understand the molecular mechanisms underlying reactivation of *Dazl* expression in our screen hits, we isolated 3 individual KO clones for each of the 4 genes of interest. All clones had loss-of-function mutations in the targeted gene (Fig. 3A and S3A), and they reactivated mScarlet to various extents: while the control cells had ~ 3% mScarlet-positive cells, *Mcm3ap* KO, *Spop* KO, and *Zbtb14* KO clones had 20–30% mScarlet-positive cells, whereas *Kdm5c* KO had ~ 45%, (Fig. 3B). The clones also expressed the untagged allele of *Dazl* at the RNA (Figure S3B) and protein level (Fig. 3C). The amount of *Dazl* expressed in serum condition by the KO clones was generally lower than the amount of *Dazl* expressed by 2i cells, showing that the induction is only suboptimal, or affects only a fraction of the cells, or both (Fig. 3C).

The ability of the phenotypes to be rescued by reintroduction of a WT cDNA was tested using PiggyBac transposons expressing V5-tagged constructs<sup>37</sup>. CRISPR-resistant alleles of mouse *Kdm5c*, *Mcm3ap*, *Spop*, and *Zbtb14* were expressed in the corresponding KO clones. In all cases, expression of the functional gene decreased the number of mScarlet-positive cells to near-background levels (Fig. 3D), rendered the cells sensitive to Hygromycin (Fig. 3E), and silenced the expression of the *Dazl* RNA (Figure S3B) and protein (Fig. 3F). These rescue experiments provide genetic evidence that loss-of-function

mutations in *Kdm5c*, *Mcm3ap*, *Spop*, and *Zbtb14* have caused the reactivation of *Dazl*, leading us to further examine the cellular phenotype of the KO cells.

The chromatin status of the mutants was characterized by Whole-Genome Bisulfite Sequencing (WGBS), at 10X median coverage (Figure S3C). A PCA analysis showed that all 4 mutants clustered together, and away from the serum-grown or 2i-grown WT cells (Figure S3D). The level of CpG methylation was reduced from 60% in WT to ~ 55% in the mutants (Fig. 3G), the difference being statistically significant for the *Mcm3ap* and *Spop* KOs. This difference was also found using orthogonal approaches, LC-MS/MS and LUMA (Figure S3E-F). The decrease affects multiple genome compartments, including enhancers and repeated elements (Fig. 3G). Nevertheless, the amount of DNA methylation decrease in the KO cells is modest, unlike that seen in 2i cells (Fig. 3G, S3E-F). Altogether, the WGBS results show that the 4 mutants have a similar alteration: a global decrease of DNA methylation of limited proportions. This experiment in itself does not distinguish between a limited decrease in all cells of the population, a larger decrease in a subset of cells, or a combination thereof. To gain further insight into the phenotype of the mutant cells, we then turned to a transcriptomic approach.

## The KOs induce a 2-cell-like transcriptional signature

mRNA-seq was performed on each of the KOs (3 individual clones for each KO) to examine the transcriptional changes present in these lines. At the thresholds we used, the KO cells showed from ~ 1,000 (*Mcm3ap* KO) to ~ 1,200 (*Zbtb14* KO) differentially expressed genes (Fig. 4A), split roughly equally between up- and down-regulated genes (Figure S4A). As expected, *Dazl* was among the upregulated genes in all mutant lines (Figs. 4A and S4B). Among 92 validated 2CLC markers<sup>15</sup>, between 21 and 32 were differentially expressed in the different KOs, including *Zscan4* and *Taf7l* (Figs. 4A and S4B). The master regulator of the 2CLC state, *Dux*, was also induced in all the KOs (Fig. 4A). The results were validated by RT-qPCR, and we verified that the genes induced in the KOs return to basal level upon rescue (Figure S4C).

The four KOs have 201 upregulated genes in common (Fig. 4B), and 2CLC markers are enriched in this common list (Fig. 4C and S4D). Correspondingly, the 201 shared upregulated genes are more expressed at the 2-cell stage than at other early developmental stages (Fig. 4D). This provides strong evidence that a 2CLC signature is present in each of the 4 KOs.

Besides the markers used above, an additional feature of 2CLC is a switch in oxidative metabolism: reduction of oxygen consumption, glycolytic activity and ROS accumulation<sup>38</sup>. GSEA analysis with Hallmark pathways showed that 'Oxidative phosphorylation', 'Glycolysis' and 'Reactive Oxygen Species' were among the top significantly depleted pathways in the combined "4KO" profile (Figure S4E), which is consistent with this switch occurring in the KO cells.

Altogether, these data show that the 4 KOs revealed by the screen reactivate a gene expression signature that is typical of 2CLCs.

## Activation of repeat elements and chimeric transcripts

After this analysis of unique genes, we then explored the expression of repeat elements (SINEs, LINEs, and ERVs) in our transcriptome data. We saw a notable induction of MERVL repeats—and of their LTRs MT2\_mm— but of few other repeats in *Kdm5c* KO, *Mcm3ap* KO, *Spop* KO, and *Zbtb14* KO clones (Fig. 4E).

A detailed analysis showed that the individual repeats activated in our 4 KOs are remarkably similar to the repeats previously reported to be upregulated in 2CLCs<sup>7,19,23</sup>, and that these same copies are also induced in the 2-cell embryo *in vivo*<sup>39</sup> (Fig. 4F). Many of these upregulated MERVL LTR copies led to the expression of chimeric transcripts, known to positively regulate the 2CLC state<sup>10</sup>. Finally, we validated our findings with RT-qPCR (Figure S4C), and also verified that ZSCAN4 and MERVL-Gag were induced at the protein level in the four KOs and silenced again when the KOs were rescued (Fig. 4G).

## Epistasis with Dux, Dppa2, and p53

We next sought to place our findings in the framework of known 2CLC regulators, and for this we used loss-of-function approaches. First, we knocked down *Dux* expression with an shRNA vector<sup>3</sup> (Fig. 5A). This caused the number of mScarlet-positive (Fig. 5B) and Hygromycin-resistant cells (Fig. 5C) to decrease to WT levels in the *Mcm3ap*, *Spop*, and *Zbtb14* KOs. In contrast, the *Kdm5c/shDux* population maintained a large fraction of mScarlet-positive and Hygromycin-resistant cells (Figs. 5B-C). We then measured the expression of selected genes by RT-qPCR in the knockdown cells (Fig. 5D). We verified that the knockdown caused a 4- to 8-fold decrease in *Dux* mRNA levels; it also caused a major decrease of 2CLC markers (*Dazl*, *Zscan4*, *Usp17le*, MERVL) in all KOs, except for the *Kdm5c* KO, in which *Dazl* expression stayed significantly higher than in WT cells (Fig. 5D). This suggests that the induction of *Dazl* is completely DUX-dependent in the *Mcm3ap*, *Spop*, and *Zbtb14* KOs, whereas it is in part DUX-independent in the *Kdm5c* KO cells.

We carried out similar experiments to examine the requirement for DPPA2 in the 2CLC induction we observed (Fig. 5E-H). Again, DPPA2 was necessary for 2CLC gene induction in all KOs, the exception being *Kdm5c* KO, in which *Dazl* induction persisted even after DPPA2 removal.

Lastly, we investigated the potential involvement of p53 (Figure S5A), as it has recently emerged as an inducer of the 2CLC state<sup>12</sup>. We used an siRNA approach and attained ~ 80–90% depletion of the *Trp53* mRNA (Figure S5B), and ~ 70% depletion of the p53 protein (Figure S5C). This had no effect on the expression of *Dazl*, *Dux*, *Usp17le* or MERVL in any of the 4 KOs (Figure S5D), showing that the *Mcm3ap*, *Spop*, *Zbtb14*, and *Kdm5c* KOs induce a 2CLC state independently of p53.

Together these results clarify the mode of action of the 4 factors under investigation. Removing MCM3AP, SPOP, or ZBTB14, induces 2CLC gene expression in a manner that is DPPA2 and DUX-dependent, showing that they act upstream of these transcriptional activators. Removal of KDM5C has a more complex, dual effect: it induces *Zscan4*, *Usp17le*, and MERVL in a DUX- and DPPA2-dependent manner, again arguing for KDM5C being an upstream regulator. However, removing KDM5C induces *Dazl* even when DUX or DPPA2 are absent, suggesting that KDM5C is repressing the *Dazl* promoter directly.

# KDM5C shares many targets with ncPRC1.6 and has a catalysis-independent role

We finally set out to investigate in more detail how the lysine demethylase KDM5C represses *Dazl* expression and inhibits the 2CLC state. In particular, we tested whether these functions were direct, and whether they were dependent on the catalytic function of KDM5C. To begin answering this question, we carried out ChIP-seq on endogenous KDM5C in DASH cells. A peak of KDM5C over the *Dazl* promoter was observed (Fig. 6A). More generally, we observed ~ 800 gene promoters bound by KDM5C, of which many belonged to germline genes and, more specifically, to germline genes that have a *Dazl*-like pattern of regulation<sup>18</sup> (Figures S6A-B). This is the first description of endogenous KDM5C distribution in mouse ESCs, and it refines previous data obtained with overexpression systems<sup>40</sup>. Genes like *Dazl* or *Taf7l*, that are both bound and repressed by KDM5C, represent only a minority of the targets (Fig. 6B), and they are highly enriched in germline genes (Figure S6C). To gain further insight into potential co-binders, we carried out a motif discovery analysis on the KDM5C peaks, and found that the highest-rated motif was the E-box bound CACGTG (Fig. 6C). This sequence is the binding site for the MAX/MGA heterodimer, a DNA binding module within ncPRC1.6, therefore suggesting that many sites directly bound by KDM5C are also bound by ncPRC1.6. The examination of MAX ChIP-seq data validated this hypothesis, as we found that 77% of KDM5C targets were also MAX targets (Fig. 6D). We conclude that KDM5C directly binds and represses a number of germline genes in ESCs, including the 2CLC markers *Dazl* and *Taf7l*. Many of these targets are also bound by ncPRC1.6 via an E-box (Fig. 6E).

This first set of data shows that most genes bound by KDM5C are also bound by ncPRC1.6, but does this overlap also extend to genes that are repressed by KDM5C even though they are not directly bound? We examined this question by comparing the transcriptomes of cells lacking *Kdm5c* or *Pcgf6* (a component of ncPRC1.6), and indeed found that there was extensive overlap: ~40% of genes repressed by KDM5C are also repressed by ncPRC1.6 (Fig. 6F), and this list includes *Dux* and other 2CLC markers (*Usp17l* cluster, *Zscan4* cluster, *Zfp352*, ...). Our ChIP-seq data did not support direct binding of KDM5C to the *Dux* locus, so we hypothesize that an indirect recruitment occurs, maybe via ncPRC1.6 (Fig. 6G).

Next we examined the requirement for the catalytic activity of KDM5C in the transcriptional repression process. *Kdm5c* KO cells rescued either with the WT version of the enzyme, or with the catalytically inactive (Cat. Inact.) mutant H514A<sup>41</sup>. As observed previously (Fig. 3), rescue with the WT form strongly reduced the number of mScarlet-positive cells emerging in the *Kdm5c* KO population. Surprisingly, the catalytically inactive mutant had the same effect (Figs. 6H). Therefore, repression of *Dazl* expression does not require lysine demethylation by KDM5C.

In addition to repressing *Dazl*, KDM5C also inhibits the 2CLC state more generally, and does so in a DUX-dependent manner (Fig. 5). The WT form of KDM5C efficiently repressed *Dux* in *Kdm5c* KO cells, as it repressed *MERVL*, *Zscan4*, and *Usp17le* (Fig. 6I). Again, catalytically inactive KDM5C repressed just as efficiently as WT (Fig. 6I), an observation we confirmed by western blotting (Fig. 6J).

To conclude, our data indicate a surprising function of KDM5C in our experimental context: catalysis-independent repression of *Dazl* and *Dux*, mediated in part by direct binding to chromatin, and in part by events that do not involve direct binding.

## Discussion

In this manuscript, we use a CRISPR KO screen in mouse ESCs to identify negative regulators of a 2CLC marker, *Dazl*. The screen yielded many of the expected hits (DNMT1, UHRF1, E2F6...), as well as a number of novel hits. A prioritization strategy led to the identification of 4 factors in which loss-of-function mutations reactivate not only *Dazl*, but a broader 2CLC signature, suggesting that they are regulators of the 2CLC status. Epigenomics, transcriptomics, and functional studies allow us to place these hits relative to known actors, and to extend the conceptual framework of 2CLC identity, with likely relevance to physiological totipotency.

## A robust screen to identify regulators of totipotency

To identify new regulators of the 2CLC state, we performed a genome-wide CRISPR KO screen, selecting for cells that reactivated the marker *Dazl*. The procedure successfully recovered new hits, along with many previously identified regulators of *Dazl*, and of the 2CLC state. We identified previously uncharacterized regulators of the 2CLC signature, there was also overlap between the hits we obtained and those found in previous approaches, including a recent CRISPR screen based on a *Dux* induction system<sup>20</sup>. Some known inhibitors of the 2CLC state failed to reach the Top 40, but are nevertheless very well ranked in our candidate list: the SUMO E3-ligase PIAS4<sup>42</sup> ranked #41; the chromatin-remodeler p400<sup>15</sup> ranked #52; the histone demethylase LSD1<sup>7</sup> ranked #111; the ncPRC1 subunits<sup>15</sup> RING1B and RYBP ranked #100 and #245 respectively. This suggests that relaxing our statistical cutoff may still yield important regulators of totipotency. In summary, our experimental scheme allowed us to fruitfully recover some known, but also some new, regulators of the 2CLC state.

One potential shortcoming of our KO approach (as opposed to siRNA or CRISPRi) is that loss of genes that are essential or near-essential is counter-selected during the selection. This likely explains why MAX, a well-validated repressor of *Dazl*<sup>16</sup>, ranks poorly in our candidate list (#1064). This may also explain why some published inhibitors of the 2CLC state, like the histone chaperone CAF-1<sup>19</sup> or the transcription factor TRIM28<sup>3</sup> did not rank high in our screen. Another constraint of the screening approach is that it uses a single reporter gene. We selected *Dazl* because it is a robust marker of the 2-cell-like stage<sup>15,20</sup>, and its transcription intensity is compatible with antibiotic selection<sup>43</sup>. This reporter yielded hits that had not been reported using other systems such as MERVL-based reporters<sup>15,20</sup>, showing its complementarity to these approaches. However, *Dazl* is also expressed in the germline, a property shared with other 2CLC markers such as *Spz1* or *Taf7l*. For this reason, we had to find ways to focus our attention on those hits that regulate the 2CLC state generally, and not just *Dazl*, or not just germline genes. A positive corollary,

however, is that our list of candidates may reveal interesting regulators of germline gene expression that could be explored in the future.

Even though these caveats exist, our reporter strain can be readily used in the future for additional screens. For instance, given the known roles of lncRNAs in the 2CLC<sup>44,45</sup>, a genome-wide CRISPRi or CRISPRoff screen on these elements may be warranted.

## Which stages of ESC to 2CLC reprogramming are affected?

The inactivation of any of our 4 hits (*Mcm3ap*, *Spop*, *Zbtb14*, and *Kdm5c*), increases the number of 2-cell-like-cells that form spontaneously in a serum culture of mouse ESCs. In theory, this could be caused by an increased rate of reprogramming of ESC to 2CLC cells, or a longer half-life of the 2CLC cells, or both (see model in Fig. 7). We have observed that *Dazl*-positive 2CLC cells sorted from the KO populations, when placed back in serum, rapidly convert back to a population that is mostly *Dazl*-negative, just like WT cells do. While this observation will require more rigorous quantitative measurements in the future, it suggests that our KOs augment the percentage of 2CLC in culture because they increase the rate of ESC-to-2CLC conversion.

The reprogramming of ESCs to 2CLCs is known to have 2 steps<sup>20</sup>: the first is the inhibition of pluripotency genes, and the second is the induction of 2C-specific/totipotency genes. Which step or steps might be regulated by the factors we have identified? Factors that impede the first step, such as *Myc*<sup>20</sup>, work by promoting the expression of pluripotency genes, such as *Sox2* and *Pou5f1*. None of our KO lines affect the expression of *Sox2* and *Pou5f1*, suggesting that the first step of ESC to 2CLC conversion is not altered. Consistent with this notion, upon spontaneous differentiation *in vitro*, the KO lines do not repress *Pou5f1* more than the WT. Therefore, we hypothesize that KDM5C, MCM3AP, SPOP and ZBTB14 repress the second stage of ESC to 2CLC reprogramming, as DNMT1 does<sup>20</sup> (Fig. 7). This hypothesis seems particularly warranted for SPOP, as its transcription is inhibited in the second step of reprogramming<sup>20</sup>. *In vivo*, these repressors of totipotency might become active immediately after the 2C stage, or later, by changes of their transcription, translation, stability, or localization (Fig. 7B).

## Transcriptional and post-transcriptional mechanisms for the action of the 4 factors

KDM5C demethylates H3K4me2/3 and is therefore a repressive demethylase<sup>41,46,47</sup>. We find that removing KDM5C from ESCs eases their conversion into 2CLCs. Intriguingly, the protein seems to have a two-pronged function. First, KDM5C directly binds and represses the promoter of certain 2CLC markers, including *Dazl* and *Taf7l*. This first observation fits with earlier findings reported in a different cell type. Indeed, KDM5C mutations in humans are linked to mental retardation, so its function has been mostly studied in the brain, and mouse neurons lacking *Kdm5c* overexpress germline genes including *Dazl*<sup>48</sup>. Second, KDM5C represses the *Dux* locus, in a catalytic-independent manner. It has been observed before that lysine demethylases can have catalysis-independent functions<sup>49</sup>, but we are not aware of any such report for KDM5C. As KDM5C has been shown to interact with E2F6 in HeLa cells<sup>46</sup>, and with PCGF6 in

mouse dendritic cells<sup>50</sup>, we hypothesize that KDM5C might be recruited by ncPRC1.6 at some of the many targets they share, including *Dux*.

Three subunits of the TREX-2 complex were recovered in our screen: MCM3AP (also known as GANP, the largest and scaffolding subunit), ENY2, and SEM1. TREX-2 has pleiotropic roles: it couples gene expression to mRNA export<sup>34-36</sup> and it also regulates the outcome of DNA repair<sup>51</sup>. Future experiments, for instance with separation-of-function mutants, will be required to determine which of these functions contribute to repressing the 2CLC state in ESCs.

SPOP is the adaptor module for an E3-ubiquitin ligase complex that promotes protein degradation<sup>31</sup>. A plausible explanation for our results would be that SPOP degrades DPPA2: indeed SPOP was found to interact with DPPA2 in a systematic interaction mapping effort<sup>52</sup>. DPPA2 directly binds the *Dux* locus and activates it<sup>10</sup>, and DPPA2 also binds to the *Dazl* promoter<sup>22</sup>. Thus, increased DPPA2 protein level caused by inactivation of *Spop* could explain the induction of *Dazl*, *Dux*, and of the 2CLC signature in the *Spop* KO cells.

Intriguingly, ATR-dependent replication stress triggers *Dux* expression<sup>53</sup>, and ZBTB14 has been recently shown to stabilize the RPA-ATR-ATRIP complex at stalled replication forks<sup>54</sup>. However, the loss of ZBTB14 leads to *decreased* ATR signaling, therefore it seems unlikely that the induction of *Dux* and of the 2CLC signature in the *Zbtb14* KO involves ATR signaling, and a direct transcriptional mechanism may be at play. Further experiments will be required to uncover this mechanism. Alternatively, given the recent discovery that replication speed controls the emergence of 2-cell-like cells<sup>55</sup>, ZBTB14 might be involved in this process.

## Relevance to totipotency, disease and therapy

Our work identifies new regulators of the 2CLC state *in vitro*. Many of the discoveries obtained using 2CLCs have proved to also apply to the 2-cell embryo<sup>1,2,8,9</sup>, therefore it is likely that MCM3AP, KDM5C, SPOP, and ZBTB14, also regulate totipotency *in vivo*. Nevertheless, the formal demonstration of this assumption will require future experiments on early mouse embryos.

Our findings have interesting consequences for human pathologies. *Dux* expression is normally restricted to a specific developmental window, and inappropriate expression of *Dux* (and its downstream targets) in skeletal muscle leads to the human disease facioscapulohumeral dystrophy (FSHD)<sup>5</sup>. Some mutations that induce *Dux* expression in ESCs seem to also induce it in muscle cells: this is at least the case for DNMT3B<sup>56</sup>, as well as TRIM28 and CAF-1<sup>57</sup>. Our results suggest that mutations altering the function of SPOP, KDM5C, MCM3AP, or ZBTB14 could potentially contribute to FSHD. Conversely, mutations of KDM5C cause mental retardation<sup>58</sup> and mutations of MCM3AP cause Charcot-Marie-Tooth disease<sup>59</sup>; it is possible that misexpression of 2CLC genes contributes to these pathologies.

Finally, in humans, the 8-cell-embryo is similar to the mouse 2-cell embryo in that it is the place of Zygotic Gene Activation, and its cells are totipotent. Very recent results have shown that human "8-cell-like" cells

can be obtained from human ESCs *in vitro*<sup>60</sup>, paving the way for further research into how these cells emerge. It will be of interest to determine whether the regulators described in our work play a conserved role in humans. This effort and those of others will help harness the potential of totipotent cells for basic research and for medicine.

## Declarations

### Accession number:

The accession number for the sequencing data reported in this paper is GEO: GSE173573.

### Author Contributions

NG and PAD conceived the project. NG, LY, and PAD planned the experiments. NG, LY, LF and AA performed experiments and analyzed the data. FM performed WGBS. LF performed LUMA. CD performed MeDIP. FB performed mass spectrometry. MD and BD performed Fluidigm. JRA performed RNA-seq and WGBS analysis. OK, ML, AS, KY, and GC, performed bioinformatic analyses. NG, LY, and PAD wrote the manuscript. PAD, TI, and NG supervised the project. GC, TI, and PAD acquired funding. All authors reviewed the manuscript.

### Competing interests statement

The authors declare no competing interests.

### Acknowledgments

We are very grateful to the following colleagues for useful advice: Allison Bardin, Brianna Rodgers, Claire Francastel, Claire Rougeulle, Sophie Polo, Pablo Navarro, Raphael Margueron, Guillaume Velasco, Guillaume Filion, Miguel Casanova, Sainitin Donakonda, Michael Weber, Laszlo Tora, Yoichi Shinkai, and Till Bartke. We thank the following colleagues for useful reagents: Déborah Bourc'his for J1 mESC, Marc Timmers for KDM5C tagged lines, Nobuo Sakaguchi for a mouse GANP cDNA, Hitoshi Niwa and Yoichi Shinkai for piggyBac constructs. We thank the Vectorology platform, Epigenetics platform, Microscopy platform and Bioinformatics/Biostatistics Core Facility (BIBS) at the CNRS Epigenetics and Cell Fate Unit (Université Paris Cité), for providing access and technical advice. We thank Emmanuelle Jeannot at Institut Curie for help with ddPCR. We warmly thank Sebastian Bultmann (LMU, Munich), for help with sgRNA sequencing and MAGECK analysis. We acknowledge the ImagoSeine core facility of the Institut Jacques Monod, member of the France Biolmaging (ANR-10-INBS-04) and the support of La ligue contre le Cancer (R03/75-79). Microfluidic RT-qPCR (Fluidigm) analysis was carried out on the qPCR-HD-Genomic Paris Centre Core Facility and was supported by grants from Région Ile-de-France, DIMBIO-RVT-INSERM-ADR-P11 n° 21016711. PAD is supported by Agence Nationale de la Recherche (PRCI INTEGER ANR-19-CE12-0030-01), LabEx "Who Am I?" (ANR-11-LABX-0071), Université de Paris IdEx (ANR-18-IDEX-0001) funded by the French Government through its "Investments for the Future" program, Fondation

pour la Recherche Médicale, Fondation ARC (Programme Labellisé PGA1/RF20180206807). PAD, AS and GC were supported by grant RETROMET, ANR-16-CE12-0020, from Agence Nationale de la Recherche. JRA and MVCG were supported by Laboratoire d'excellence Who Am I? (Labex 11-LABX-0071) Emerging Teams Grant and by the European Research Council (ERC-StG-2019 DyNamecs). This research was supported by Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number JP20am0101103 (support number 2652). KY was the recipient of a postdoctoral fellowship from Fondation Association pour la Recherche sur le Cancer, and of a subsequent postdoc fellowship from Labex WhoAml. ML thanks the Ligue contre le Cancer for a 4th year PhD fellowship.

We apologize to authors whose primary references we could not cite because of space limitations.

## Materials And Methods

### *Cell culture*

J1 mouse ESCs (129S4/SvJae, XY) were cultured on gelatin-coated dishes in serum/LIF medium containing DMEM/GlutaMAX supplemented with 15% fetal bovine serum (FBS), non-essential amino acids (NEAA), penicillin/streptomycin, and 1000 U/ml leukemia inhibitory factor (LIF). When necessary, ESCs were adapted to 2i/Vitamin C/LIF medium containing serum-free DMEM-F12 and Neurobasal media supplemented with 1% N2, 2% B27, 100 µg/ml ascorbic acid, 1 µM PD0325901, and 3 µM CHIR99021. Cells were incubated in a humidified atmosphere at 37°C under 5% CO<sub>2</sub>. For spontaneous differentiation assays, cells were seeded at clonal density in a serum medium without LIF.

### *Cloning of sgRNA, transfection, and transduction in ESCs*

The single-guide RNAs (sgRNAs) were designed using either Benchling/CRISPOR software or sequences were obtained from the Brie library. The sgRNAs were cloned either in the PX459 vector (Addgene #62988) or in lentiCRISPRv2 (Addgene #52961). For ESC transfection, we used an Amaxa 4D-Nucleofector (Lonza), according to the manufacturer's instructions. Production of lentiviral particles was performed by calcium-phosphate transfection of HEK293T with psPAX2 and pMD2.G plasmids, in a BSL3 tissue culture facility.

### *Generation of DASH (Dazl-mScarlet-Hygromycin<sup>R</sup>) reporter cell line*

The reporter cassette P2A-mScarlet-T2A-Hygro<sup>R</sup> (synthesized by GenScript) was inserted in-frame within the exon 6 of mouse *Dazl* gene. The cassette was flanked by *Dazl* homology arms (HA) corresponding to endogenous intron 5-exon 6 and intron 6 sequences respectively. Protospacer Adjacent Motif (PAM) sites of the two sgRNAs targeting *Dazl* exon 6 were mutated in the homology arms to prevent re-cutting of Cas9 after the insertion of the cassette through homologous recombination. The synthesized cassette was cloned into pUC57-Simple. The two sgRNAs targeting *Dazl* exon 6 were cloned into the

pSpCas9(BB)-2A-GFP backbone (Addgene #48138). The homologous integration of the reporter cassette in one of the alleles of *Dazl* was confirmed by PCR and sequencing.

### ***CRISPR KO screen: amplification of sgRNA library, lentiviral transduction, and sample collection***

We performed a genome-wide CRISPR knock-out (KO) screen using the lentiviral Brie sgRNA library containing 4 sgRNAs per protein-coding gene (Addgene #73632)<sup>27</sup>. The sgRNA library was amplified and sequenced confirming an equal representation of the ~80,000 sgRNAs, as expected and lentiviruses were produced. The screen was performed in two biological replicates, cells were transduced with a lentiviral plasmid pool at a multiplicity of infection (MOI) ~0.1. After 48 h, transduced cells were positively selected with 2 µg/ml puromycin for 5 days. Coverage was 150X (150 transduced cells/sgRNA) for each biological replicate. Sequencing of the Puro<sup>R</sup> sample confirmed a comprehensive representation of the sgRNA library in the screen. Following this, cells were initially selected with 50 µg/ml Hygromycin for 3 days followed by additional selection for 11 days at 125 µg/ml Hygromycin. Three weeks post-infection, Hygromycin-resistant cells were sorted by FACS for mScarlet expression.

### ***CRISPR KO screen: sequencing and analysis***

Genomic DNA was isolated from cells using AllPrep DNA/RNA Mini Kit (Qiagen) following the recommended protocol. For the sgRNA plasmid library, 50 ng DNA was used for each reaction whereas 300 ng of genomic DNA was used per reaction. Multiple reactions were set-up for each sample to reflect the coverage and PCR reactions were pooled. Briefly, PCR was performed with Platinum Taq polymerase (Thermo Fisher Scientific), employing a pool of P5 primers (designed in the lab) and a unique P7 barcode primer (primer sequences are listed in Supplementary File 1). The PCR conditions were: initial denaturation at 94°C for 4 min; 28 cycles of denaturation at 94°C for 30 s, annealing at 53°C for 30 s and extension at 72°C for 30 s per kb; final extension at 72°C for 10 min. The PCR products were retrieved using QIAquick PCR Purification Kit (Qiagen) and verified for the right amplicon on an agarose gel. The DNA was further purified using AMPure XP (Beckman Coulter). Libraries were sequenced on an Illumina HiSeq 1500 in single-end (SE) 100 bp output mode. The sgRNA distribution and enrichment at different time points were analyzed with the MAGeCK workflow<sup>28</sup>. A statistical threshold of p-value < 0.0005 resulted in a list of 40 candidates whose knockout led to the expression of mScarlet and Hygro<sup>R</sup> in DASH ESCs. The list of enriched genes is available in Supplementary File 2.

### ***CRISPR KO screen: protein-protein interaction network analysis***

Protein-protein interactions for the 40 significant hits from the genome-wide CRISPR KO screen were performed using the STRING v11.0 tool. Interactions were computed using default parameters and network edges with a confidence score > 0.6 were displayed. Network visualization was edited with Cytoscape software.

### ***Generation of individual gene KO and clonal cell lines***

To validate candidates from the screen, individual KOs in DASH ESCs were generated using the top two most efficient sgRNAs (as determined by MAGeCK analysis). During lentiviral production, both sgRNA plasmids targeting the same gene were mixed to increase knockout efficiency. Transduced cells were selected with 2 µg/ml puromycin for 3 days; followed by Hygromycin selection (50 µg/ml for 3 days, and 125 µg/ml for the next 7 days). For *Kdm5c*, *Mcm3ap*, *Spop*, and *Zbtb14*, three independent clonal KO lines were established.

### ***Rescue experiments, piggyBac system***

For rescue experiments, the coding sequence (CDS) of candidates was either synthesized (*Spop* and *Zbtb14*) (GenScript), amplified from cDNA (*Kdm5c*), or obtained from colleagues (*Mcm3ap*, kindly shared by Prof. N. Sakaguchi<sup>32</sup>). In all cases, silent mutations were incorporated either within the PAM and/or sgRNA sequence to block targeting by the active cognate sgRNAs in the KO clonal cell lines. For *Kdm5c*, an additional catalytically inactive mutant plasmid with the H514A mutation was generated. These CDS sequences were cloned into a piggyBac vector and co-transfected with PB transposase for stable insertion<sup>36</sup>. Empty piggyBac vector served as a control. Transfected cells were selected with 5 µg/ml blasticidin for 5 days and processed for phenotypic and molecular assays.

### ***Knock-down experiments, siRNA/shRNA transfection***

For stable knock-down experiments, the pLKO.1-blasticidin shRNA vector for *Dux* was kindly shared by Prof. Didier Trono<sup>3</sup>, and the pLKO.1-neomycin shRNA vector for *Dppa2* was ordered from Sigma (TRCN0000174599). In both instances, empty backbones were used as controls. Transduced cells were respectively selected with 5 µg/ml blasticidin for 5 days or 400 µg/ml geneticin for 7 days, then processed for phenotypic and molecular assays. For transient knock-down experiments, the siRNA vector for *Tp53* was kindly shared by G. Velasco<sup>26</sup>. A non-targeting siRNA pool (siGenome siControl Pool #2, Dharmacon) was used as a control. Cells were transfected with 50 nM siRNA and 3 µl/ml Lipofectamine RNAiMAX (Thermo Fisher Scientific) diluted in Opti-MEM. Total RNA was extracted 2 days after transfection for RT-qPCR. Primer sequences are listed in Supplementary File 1.

### ***Flow cytometry***

The number of cells expressing mScarlet was determined by flow cytometry using a yellow laser (561 nm) of the Influx or FACSAria Fusion cell sorter (BD Biosciences) at the ImagoSeine core facility (Institut Jacques Monod). A threshold on mScarlet signal intensity (subtracting background fluorescence from non-reporter wild-type mESC) was used to determine the proportion of positive cells. Data were analyzed with FlowJo software.

### ***Crystal violet staining***

Cells were seeded at the same density in all wells and grown with or without Hygromycin for 7 days. Surviving cells were fixed with absolute ethanol for 15 min, stained with 1% Crystal violet dye (Sigma) for

30 min, and washed extensively with water to remove the unbound stain.

### ***Digital droplet PCR (ddPCR)***

The PCR reaction mixture composed of 2X EvaGreen ddPCR Supermix (Bio-Rad), primers at a final concentration of 100 nM and 10 ng of template DNA were partitioned into up to 20,000 droplets by water-oil emulsion. After droplet generation, a regular PCR was performed with the following conditions: 95°C for 5 min; 95°C for 30 s and 60°C for 1 min (40 cycles); 4°C for 5 min, 90°C for 5 min, 4°C hold. For all steps, a ramp rate of 2°C/s was used. Cycled droplets were read individually (Bio-Rad QX-200 droplet reader). Each run included technical duplicates and no-template controls. Primer sequences are listed in Supplementary File 1.

### ***RT-qPCR***

Total RNA was extracted from cells with RNeasy Plus Mini kit (Qiagen) according to the manufacturer's instructions and quantified using Qubit RNA BR Assay kit on Qubit 2.0 Fluorometer (Thermo Fisher Scientific). One microgram of total RNA was reverse transcribed using SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific) and Oligo dT primers (Promega). RT-qPCR was performed using Power SYBR Green (Applied Biosystems) on a ViiA 7 Real-Time PCR System (Thermo Fisher Scientific). *Actinb*, *Ppia*, and *Rplp0* were used for normalization. Primer sequences are available in Supplementary File 1.

### ***Western blotting***

Cells were harvested and lysed in RIPA buffer (Sigma) with protease inhibitor cocktail (Thermo Fisher Scientific), sonicated with a series of 30s ON / 30s OFF for 5 min on a Bioruptor (Diagenode), and centrifuged at 16,000 g for 5 min at 4°C. The supernatant was collected and quantified by BCA assay (Thermo Fisher Scientific). Thirty microgram protein extract per sample was mixed with NuPage 4X LDS Sample Buffer and 10X Sample Reducing Agent (Thermo Fisher Scientific) and denatured at 95°C for 5 min. Samples were resolved on a pre-cast SDS-PAGE 4-12% gradient gel (Thermo Fisher Scientific) with 120V electrophoresis for 90 min and blotted onto a nitrocellulose membrane (Millipore). The membrane was blocked with 5% fat-free milk/PBS at RT for 1 h, then incubated overnight at 4°C with appropriate primary antibodies. After three washes with PBS/0.1% Tween20, the membranes were incubated with the cognate fluorescent secondary antibodies and revealed in the LI-COR Odyssey-Fc imaging system. The following antibodies were used in this study: α-DAZL (Abcam #ab34139, 1:500), α-V5 (Abcam #ab206566, 1:1000), α-MuERVL-Gag (HuaBio #ER50102, 1:1000), α-ZSCAN4 (Merck #AB4340, 1:5000), α-p53 (CST #2524; 1:1000), α-TUBULIN (Abcam #7291; 1:10000), α-GAPDH (Abcam #ab9485, 1:10000). The following secondary antibodies were used in this study: IRDye 800CW Donkey α-Rabbit (Licor #926-32213, 1:15000) and IRDye 680RD Donkey α-Mouse (Licor #926-68072, 1:15000).

### ***Chromatin Immunoprecipitation (ChIP)***

$1 \times 10^7$  cells were cross-linked with 1% formaldehyde for 5 min at room temperature. Cross-linking was stopped by adding Glycine (125 mM final), and the cells were washed with PBS. Cells were resuspended in Swelling Buffer (0.5% NP-40, 0.85 mM KCl, 1 mM PMSF, 5 mM PIPES pH 8.0) supplemented with protease inhibitor cocktail (Roche) and incubated for 20 min on ice. After centrifugation, cell nuclei were resuspended in IP Buffer (0.1% SDS, 1% Triton X-100, 3 mM EDTA, 1 mM PMSF, 150 mM NaCl, 25 mM Tris-HCl pH 8.0, 1X protease inhibitor cocktail) and sonicated for 5 min (series of 30s ON / 30s OFF) on a Bioruptor Pico (Diagenode) to generate 200 to 500 bp fragments. Fragmented chromatin (50  $\mu$ g) was immunoprecipitated in IP buffer with 1  $\mu$ g of antibody (a-KDM5C, Bethyl Laboratories #A301-034A) overnight at 4°C. Subsequent steps (including incubation with magnetic beads, multiple washes, elution) were performed with the Pierce Magnetic ChIP Kit (Thermo Fisher Scientific), following the manufacturer's instructions. Reverse cross-linked DNA was purified by ChIP DNA Clean&Concentrator kit (Zymo Research). Libraries were prepared with KAPA HyperPrep kit (Roche), following the manufacturer's instructions.

### ***ChIP-seq analysis***

FASTQ reads were trimmed using Trimmomatic (v0.39) and parameters: ILLUMINACLIP:illumina\_adapters.fa:2:30:10 SLIDINGWINDOW:4:20 MINLEN:36. Trimmed reads were aligned using Bowtie2 (v2.4.1) in `-local` mode. Following alignment, Picard (v2.23.4) CleanSam, SamFormatConverter, SortSam and MarkDuplicates were used to generate a duplicate-marked bam file. The resulting bam files were converted to bigwig using deeptools (v3.3.0) Bamcoverage and options `-ignoreDuplicates -normalizeUsing CPM -minMappingQuality 10 -ignoreForNormalization chrX chrY chrM`. Peaks were called using MACS2 with default parameters, and motif enrichment analysis was performed with HOMER. For datasets already published, SRA files were downloaded from NCBI GEO using E-utilities `esearch` and `efetch` (v15.9) and converted to FASTQ using SRAToolkit (v2.8.0). The datasets generated in this study are listed in Supplementary File 3.

### ***Isolation of genomic DNA***

Genomic DNA was isolated from cells using overnight 200  $\mu$ g/ml proteinase K treatment at 55°C followed by 20  $\mu$ g/ml RNase A treatment at 37°C for 1 h and extracted by standard phenol/chloroform/alcohol method. Alternatively, genomic DNA was isolated from cells using QIAamp DNA Mini kit (Qiagen), following the manufacturer's instructions. Genomic DNA was resuspended in water and quantified with Qubit dsDNA BR Assay kit on Qubit 2.0 Fluorometer (Thermo Fisher Scientific). DNA integrity was assessed with Genomic DNA ScreenTape on TapeStation system (Agilent) and samples with DNA Integrity Number > 9 were used for subsequent analysis.

### ***DNA methylation analysis: Methylated DNA Immunoprecipitation (MeDIP)***

MeDIP was performed using the Auto MeDIP Kit on an automated platform SX-8G IP-Star Compact (Diagenode). Briefly, 2.5  $\mu$ g of DNA was sheared using a Bioruptor Pico to approximately 500-bp fragments, as assessed with D5000 ScreenTape (Agilent). Cycle conditions were as follows: 15 s ON / 90

s OFF, repeated 6 times. A portion of sheared DNA (10%) was kept as input and the rest of the sheared DNA was immunoprecipitated with  $\alpha$ -5-methylcytosine antibody (Diagenode), bound to magnetic beads, and was isolated. qPCR for selected genomic loci was performed and efficiency was calculated as % (me-DNA-IP/total input). Primer sequences are listed in Supplementary File 1.

### ***DNA methylation analysis: Luminometric methylation assay (LUMA)***

To assess global CpG methylation, 500 ng of genomic DNA was digested with MspI+EcoRI and HpaII+EcoRI (NEB) in parallel reactions, EcoRI was included as an internal reference. CpG methylation percentage is defined as the HpaII/MspI ratio. Samples were analyzed using PyroMark Q24 Advanced pyrosequencer

### ***DNA methylation analysis: LC-MS/MS***

The genomic DNA was extracted as described above with an additional step of digestion with RNase A. One microgram of DNA was treated with 10U DNA Degradase Plus (ZymoResearch) at 37°C for 4 h. After enzyme inactivation at 70°C for 20 min, the solution was filtered with Amicon Ultra-0.5 mL 10 K centrifugal filters (Merck Millipore). The reaction mix retained in the centrifugal filter was processed for LC-MS/MS analysis. Analysis of global levels of 5-mdC were performed on a Q exactive mass spectrometer (Thermo Fisher Scientific). It was equipped with an electrospray ionization source (H-ESI II Probe) coupled with an Ultimate 3000 RS HPLC (Thermo Fisher Scientific). Digested DNA was injected onto a ThermoFisher Hypersil Gold aQ chromatography column (100 mm \* 2.1 mm, 1.9  $\mu$ m particle size) heated at 30°C. The flow rate was set at 0.3 ml/min and run with an isocratic eluent of 1% acetonitrile in water with 0.1% formic acid for 10 minutes. Parent ions were fragmented in positive ion mode with 10% normalized collision energy in parallel-reaction monitoring (PRM) mode. MS2 resolution was 17,500 with an AGC target of 2e5, a maximum injection time of 50 ms, and an isolation window of 1.0 m/z. The inclusion list contained the following masses: dC (228.1) and 5-mdC (242.1). Extracted ion chromatograms of base fragments ( $\pm$ 5ppm) were used for detection and quantification (112.0506 Da for dC; 126.0662 Da for 5-mdC). Calibration curves were previously generated using synthetic standards in the ranges of 0.2 to 10 pmol injected for dC and 0.02 to 10 pmol for 5mdC. Results are expressed as a percentage of total dC.

### ***Whole-genome-bisulfite sequencing (WGBS)***

Genomic DNA was extracted as described for LC-MS/MS. The library preparation for WGBS was performed with the tPBAT protocol described previously<sup>59,60</sup>. One hundred nanograms of genomic DNA spiked with 1% (w/w) of unmethylated lambda DNA (Promega) was used for the library preparation. The sequencing was performed by MacroGen Japan Inc. using the HiSeq X Ten system. We assigned 8 lanes for the analysis of 20 samples. The sequenced reads were mapped with BMap and summarized with an in-house pipeline as described previously<sup>60</sup>, with custom scripts archived using GitHub (<https://github.com/FumihitoMiura/Project-2>). The basic metrics of the methylome data are provided in Supplementary File 4. DNA methylation levels over CpGs covered by at least 5 sequencing reads were

averaged over the following regions of interest: genome-wide 2kb bins, enhancer elements (Ensembl Regulatory features release 81, n=73,796), promoters (NCBI Refseq, TSS +/- 1kb, n=24,371), gene bodies and transposable elements (RepeatMaster, n=5,147,736). PCA plot was generated using Deeptools multiBamSummary with default parameters.

### ***RNA-seq: library preparation for transcriptome sequencing***

A total amount of 1 µg total RNA per sample was used as input material for the RNA sample preparations. RNA samples were spiked with ERCC RNA Spike-In Mix (Thermo Fisher Scientific). Sequencing libraries were generated using NEBNext Ultra™ RNA Library Prep Kit for Illumina (NEB) following the manufacturer's recommendations. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer (5X). First-strand cDNA was synthesized using a random hexamer primer and M-MuLV Reverse Transcriptase (RNase H-). Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. In the reaction buffer, dNTPs with dTTP were replaced by dUTP. The remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3' ends of DNA fragments, NEBNext Adaptor with hairpin loop structure was ligated to prepare for hybridization. To select cDNA fragments of preferentially 250-300 bp in length, the library fragments were purified with the AMPure XP system (Beckman Coulter). Then 3 µl USER Enzyme (NEB) was used with size-selected, adaptor-ligated cDNA at 37°C for 15 min followed by 5 min at 95°C before PCR. Then PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers, and Index (X) Primer. At last, products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system.

### ***RNA-seq: read alignment***

FASTQ reads were trimmed using Trimmomatic (v0.39) and parameters: ILLUMINACLIP:adapters.fa:2:30:10 SLIDINGWINDOW:4:20 MINLEN:36. Read pairs that survived trimming were aligned to the mouse reference genome (build mm10) using STAR (v2.7.5c) and default single-pass parameters. PCR duplicate read alignments were flagged using Picard-tools (2019) MarkDuplicates (v2.23.4). Uniquely aligned, non-PCR-duplicate reads were kept for downstream analysis using Samtools view (v1.10) and parameters: -q 255 -F 1540. Gene expression values were calculated over the mm10 NCBI RefSeq Genes annotation using VisRseq (v0.9.12) and normalized per million aligned reads per transcript length in kilobases (RPKM). Bigwig files were generated using deeptools bamCoverage (v3.3.0) using counts per million (CPM) normalization and visualized in IGV (v2.8.9).

### ***RNA-seq: differential expression and gene-set enrichment analysis***

DESeq2 (v1.30.0) was employed using apeglm LFC shrinkage to calculate differential expression. Genes or transposable elements were categorized as significantly differentially expressed if they showed an absolute expression fold-change  $\geq 2$  and associated adjusted p-value  $< 0.01$ . Differentially expressed genes are listed in Supplementary File 5. Gene set enrichment analysis was performed using GSEA

(v4.1.0) and default parameters (1000 permutations, permutation type = gene\_set. Selected significant terms from Hallmark gene sets (n=50 “h.all.v7.4.symbols.gmt”) were displayed. Complete GSEA analysis is available in Supplementary File 6.

### ***RNA-seq: transposable element quantification***

RepeatMasker (last updated 2012-02-06) was downloaded from the UCSC Table Browser. To measure the expression of transposable element families, PCR duplicates were removed and all reads, including uniquely mapped and multi-mapped reads, were enumerated using VisRseq. Multi-mapped reads were counted once, and all individual elements were aggregated to calculate family-wide expression in read count for differential expression analysis. Heatmaps were generated using Morpheus (<https://software.broadinstitute.org/morpheus>).

### ***RNA-sequencing analysis: MERVL/MT2\_Mm analysis***

MERVL internal sequences and their MT2\_Mm LTR promoters were extracted from the RepeatMasker annotation (last updated 2012-02-06). Internal sequences and their LTRs located within 88 bp were merged into a single element using bedtools merge (v2.27.0) to account for an 87 bp insertion of a related ORR1A3 element. Elements were categorized as full-length MERVL elements if they contained both LTR elements and internal sequences and spanned >6000bp. MT2\_Mm elements under 500 bp in length were defined as “Solo-LTRs”. All other elements, such as those composed of MERVL internal sequences and only one LTR, were categorized as “other”. Genome-wide mappability scores were calculated using iGEM (v1.315) and parameters: K\_MER\_SIZE=300 MAX\_MISMATCHES=0.04 and the mappability of each MERVL element was calculated using VisRseq. A list of MERVL elements that generate chimeric transcripts was downloaded<sup>7</sup> and mapped onto the mm10 genome using UCSC LiftOver. To measure individual transposable element expression, only uniquely aligned, non-PCR duplicate reads were counted. Elements were grouped and sorted by K-medoid clustering on log10-transformed RPKM values using the R package “cluster” and VisRseq.

## **References**

1. Ladstätter, S. & Tachibana, K. Genomic insights into chromatin reprogramming to totipotency in embryos. *J Cell Biol* **218**, 70–82 (2019).
2. Riveiro, A. R. & Brickman, J. M. From pluripotency to totipotency: an experimentalist’s guide to cellular potency. *Development* **147**, dev189845 (2020).
3. De Iaco, A. *et al.* DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat Genet* **49**, 941–945 (2017).
4. Hendrickson, P. G. *et al.* Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat Genet* **49**, 925–934 (2017).
5. Whiddon, J. L., Langford, A. T., Wong, C.-J., Zhong, J. W. & Tapscott, S. J. Conservation and innovation in the DUX4-family gene network. *Nature Genetics* **49**, 935–940 (2017).

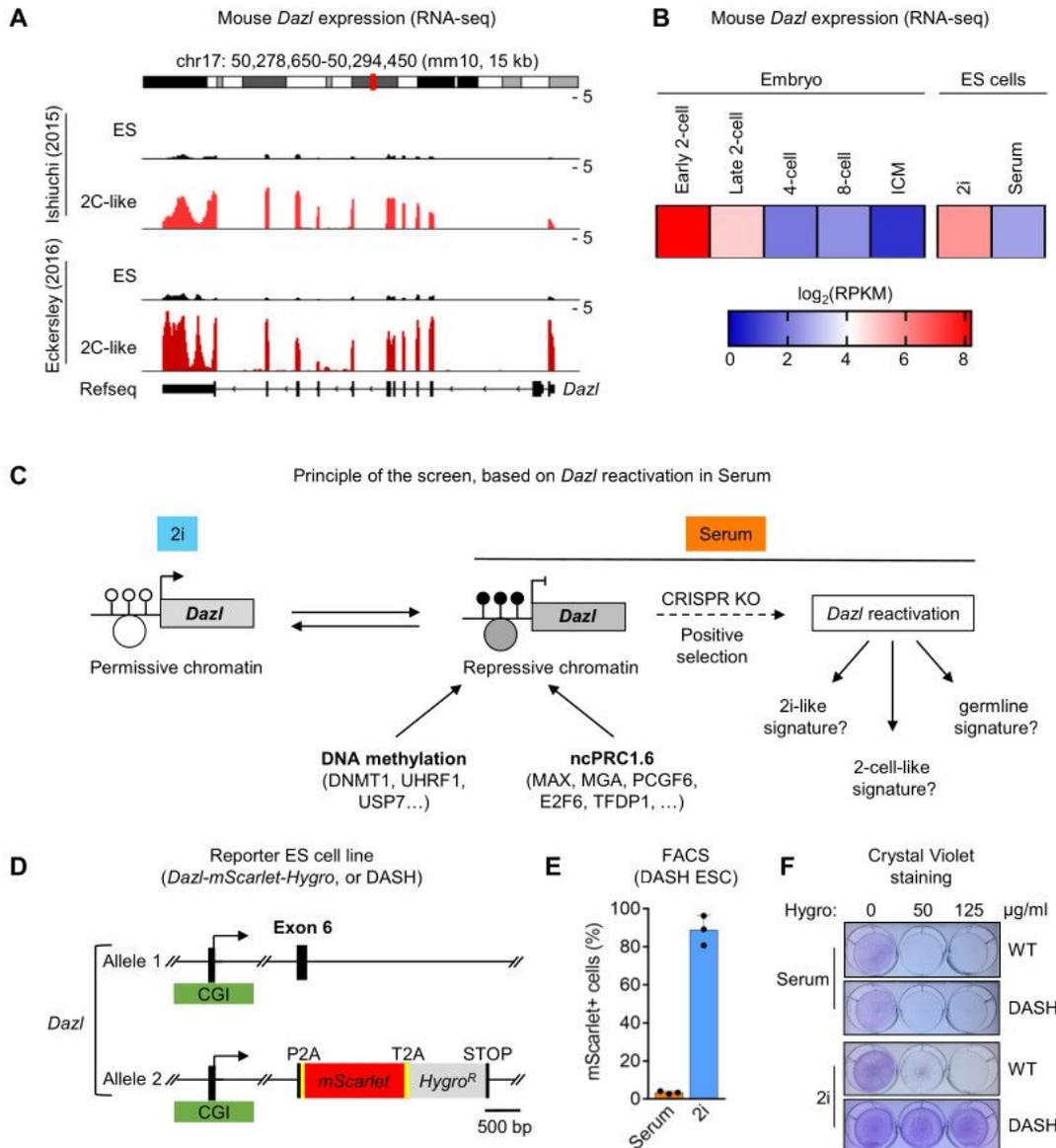
6. Falco, G. *et al.* Zscan4: a novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Dev Biol* **307**, 539–550 (2007).
7. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487**, 57–63 (2012).
8. Genet, M. & Torres-Padilla, M.-E. The molecular and cellular features of 2-cell-like cells: a reference guide. *Development* **147**, (2020).
9. Iturbide, A. & Torres-Padilla, M.-E. A cell in hand is worth two in the embryo: recent advances in 2-cell like cell reprogramming. *Current Opinion in Genetics & Development* **64**, 26–30 (2020).
10. Eckersley-Maslin, M. *et al.* Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. *Genes Dev.* **33**, 194–208 (2019).
11. De Iaco, A., Coudray, A., Duc, J. & Trono, D. DPPA2 and DPPA4 are necessary to establish a 2C-like state in mouse embryonic stem cells. *EMBO Rep* **20**, e47382 (2019).
12. Grow, E. J. *et al.* p53 convergently activates Dux/DUX4 in embryonic stem cells and in facioscapulohumeral muscular dystrophy cell models. *Nat Genet* **53**, 1207–1220 (2021).
13. Yang, M. *et al.* Chemical-induced chromatin remodeling reprograms mouse ESCs to totipotent-like stem cells. *Cell Stem Cell* S1934590922000108 (2022) doi:10.1016/j.stem.2022.01.010.
14. Morgan, M., Kumar, L., Li, Y. & Baptissart, M. Post-transcriptional regulation in spermatogenesis: all RNA pathways lead to healthy sperm. *Cell Mol Life Sci* **78**, 8049–8071 (2021).
15. Rodriguez-Terrones, D. *et al.* A molecular roadmap for the emergence of early-embryonic-like cells in culture. *Nat Genet* **50**, 106–119 (2018).
16. Endoh, M. *et al.* PCGF6-PRC1 suppresses premature differentiation of mouse embryonic stem cells by regulating germ cell-related genes. *Elife* **6**, (2017).
17. Dahlet, T. *et al.* E2F6 initiates stable epigenetic silencing of germline genes during embryonic development. *Nat Commun* **12**, 3582 (2021).
18. Mochizuki, K. *et al.* Repression of germline genes by PRC1.6 and SETDB1 in the early embryo precedes DNA methylation-mediated silencing. *Nat Commun* **12**, 7020 (2021).
19. Ishiuchi, T. *et al.* Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. *Nat Struct Mol Biol* **22**, 662–671 (2015).
20. Fu, X., Wu, X., Djekidel, M. N. & Zhang, Y. Myc and Dnmt1 impede the pluripotent to totipotent state transition in embryonic stem cells. *Nat Cell Biol* **21**, 835–844 (2019).
21. Alda-Catalinas, C. *et al.* A Single-Cell Transcriptomics CRISPR-Activation Screen Identifies Epigenetic Regulators of the Zygotic Genome Activation Program. *Cell Systems* **11**, 25–41.e9 (2020).
22. Gretarsson, K. H. & Hackett, J. A. Dppa2 and Dppa4 counteract de novo methylation to establish a permissive epigenome for development. *Nat Struct Mol Biol* (2020) doi:10.1038/s41594-020-0445-1.
23. Eckersley-Maslin, M. A. *et al.* MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. *Cell Reports* **17**, 179–192 (2016).

24. Habibi, E. *et al.* Whole-Genome Bisulfite Sequencing of Two Distinct Interconvertible DNA Methylomes of Mouse Embryonic Stem Cells. *Cell Stem Cell* **13**, 360–369 (2013).
25. Leitch, H. G. *et al.* Naive pluripotency is associated with global DNA hypomethylation. *Nat Struct Mol Biol* **20**, 311–316 (2013).
26. Velasco, G. *et al.* Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine somatic tissues. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9281–9286 (2010).
27. Laisné, M., Gupta, N., Kirsh, O., Pradhan, S. & Defossez, P.-A. Mechanisms of DNA Methyltransferase Recruitment in Mammals. *Genes (Basel)* **9**, (2018).
28. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* **34**, 184–191 (2016).
29. Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol* **15**, 554 (2014).
30. Petryk, N., Bultmann, S., Bartke, T. & Defossez, P.-A. Staying true to yourself: mechanisms of DNA methylation maintenance in mammals. *Nucleic Acids Res* (2020) doi:10.1093/nar/gkaa1154.
31. Clark, A. & Burleson, M. SPOP and cancer: a systematic review. *Am J Cancer Res* **10**, 704–726 (2020).
32. Li, M. *et al.* Genome-wide CRISPR-KO Screen Uncovers mTORC1-Mediated Gsk3 Regulation in Naive Pluripotency Maintenance and Dissolution. *Cell Reports* **24**, 489–502 (2018).
33. Singh, S. K. *et al.* GANP regulates recruitment of AID to immunoglobulin variable regions by modulating transcription and nucleosome occupancy. *Nat Commun* **4**, 1830 (2013).
34. Aksenova, V. *et al.* Nucleoporin TPR is an integral component of the TREX-2 mRNA export pathway. *Nat Commun* **11**, 4577 (2020).
35. Umlauf, D. *et al.* The human TREX-2 complex is stably associated with the nuclear pore basket. *Journal of Cell Science* **126**, 2656–2667 (2013).
36. Schneider, M. *et al.* The Nuclear Pore-Associated TREX-2 Complex Employs Mediator to Regulate Gene Expression. *Cell* **162**, 1016–1028 (2015).
37. Fukuda, K., Okuda, A., Yusa, K. & Shinkai, Y. A CRISPR knockout screen identifies SETDB1-target retroelement silencing factors in embryonic stem cells. *Genome Res.* **28**, 846–858 (2018).
38. Rodriguez-Terrones, D. *et al.* A distinct metabolic state arises during the emergence of 2-cell-like cells. *EMBO Rep* **21**, e48354 (2020).
39. Zhang, W. *et al.* Zscan4c activates endogenous retrovirus MERVL and cleavage embryo genes. *Nucleic Acids Research* gkz594 (2019) doi:10.1093/nar/gkz594.
40. Outchkourov, N. S. *et al.* Balancing of Histone H3K4 Methylation States by the Kdm5c/SMCX Histone Demethylase Modulates Promoter and Enhancer Function. *Cell Reports* **3**, 1071–1079 (2013).
41. Iwase, S. *et al.* The X-linked mental retardation gene SMCX/JARID1C defines a family of histone H3 lysine 4 demethylases. *Cell* **128**, 1077–1088 (2007).

42. Yan, Y.-L. *et al.* DPPA2/4 and SUMO E3 ligase PIAS4 opposingly regulate zygotic transcriptional program. *PLOS Biology* **17**, e3000324 (2019).
43. Nakatake, Y. *et al.* Kinetics of drug selection systems in mouse embryonic stem cells. *BMC Biotechnol* **13**, 64 (2013).
44. Percharde, M. *et al.* A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* **174**, 391–405.e19 (2018).
45. Jachowicz, J. W. *et al.* LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat Genet* **49**, 1502–1510 (2017).
46. Tahiliani, M. *et al.* The histone H3K4 demethylase SMCX links REST target genes to X-linked mental retardation. *Nature* **447**, 601–605 (2007).
47. Christensen, J. *et al.* RBP2 belongs to a family of demethylases, specific for tri- and dimethylated lysine 4 on histone 3. *Cell* **128**, 1063–1076 (2007).
48. Scandaglia, M. *et al.* Loss of Kdm5c Causes Spurious Transcription and Prevents the Fine-Tuning of Activity-Regulated Enhancers in Neurons. *Cell Reports* **21**, 47–59 (2017).
49. Shpargel, K. B., Sengoku, T., Yokoyama, S. & Magnuson, T. UTX and UTY Demonstrate Histone Demethylase-Independent Function in Mouse Embryonic Development. *PLOS Genetics* **8**, e1002964 (2012).
50. Boukhaled, G. M. *et al.* The Transcriptional Repressor Polycomb Group Factor 6, PCGF6, Negatively Regulates Dendritic Cell Activation and Promotes Quiescence. *Cell Rep* **16**, 1829–1837 (2016).
51. Evangelista, F. M. *et al.* Transcription and mRNA export machineries SAGA and TREX-2 maintain monoubiquitinated H2B balance required for DNA repair. *Journal of Cell Biology* **217**, 3382–3397 (2018).
52. Rolland, T. *et al.* A Proteome-Scale Map of the Human Interactome Network. *Cell* **159**, 1212–1226 (2014).
53. Atashpaz, S. *et al.* ATR expands embryonic stem cell fate potential in response to replication stress. *eLife* **9**, e54756 (2020).
54. Kim, W. *et al.* ZFP161 regulates replication fork stability and maintenance of genomic stability by recruiting the ATR/ATRIP complex. *Nature Communications* **10**, 5304 (2019).
55. Nakatani, T. *et al.* DNA replication fork speed underlies cell fate changes and promotes reprogramming. *Nat Genet* **54**, 318–327 (2022).
56. van den Boogaard, M. L. *et al.* Mutations in DNMT3B Modify Epigenetic Repression of the D4Z4 Repeat and the Penetrance of Facioscapulohumeral Dystrophy. *The American Journal of Human Genetics* **98**, 1020–1029 (2016).
57. Campbell, A. E. *et al.* NuRD and CAF-1-mediated silencing of the D4Z4 array is modulated by DUX4-induced MBD3L proteins. *eLife* **7**, e31023 (2018).
58. Bonefas, K. M. & Iwase, S. Soma-to-germline transformation in chromatin-linked neurodevelopmental disorders? *FEBS J* (2021) doi:10.1111/febs.16196.

59. Ylikallio, E. *et al.* MCM3AP in recessive Charcot-Marie-Tooth neuropathy and mild intellectual disability. *Brain* **140**, 2093–2103 (2017).
60. Taubenschmid-Stowers, J. *et al.* 8C-like cells capture the human zygotic genome activation program in vitro. *Cell Stem Cell* S1934-5909(22)00050–9 (2022) doi:10.1016/j.stem.2022.01.014.

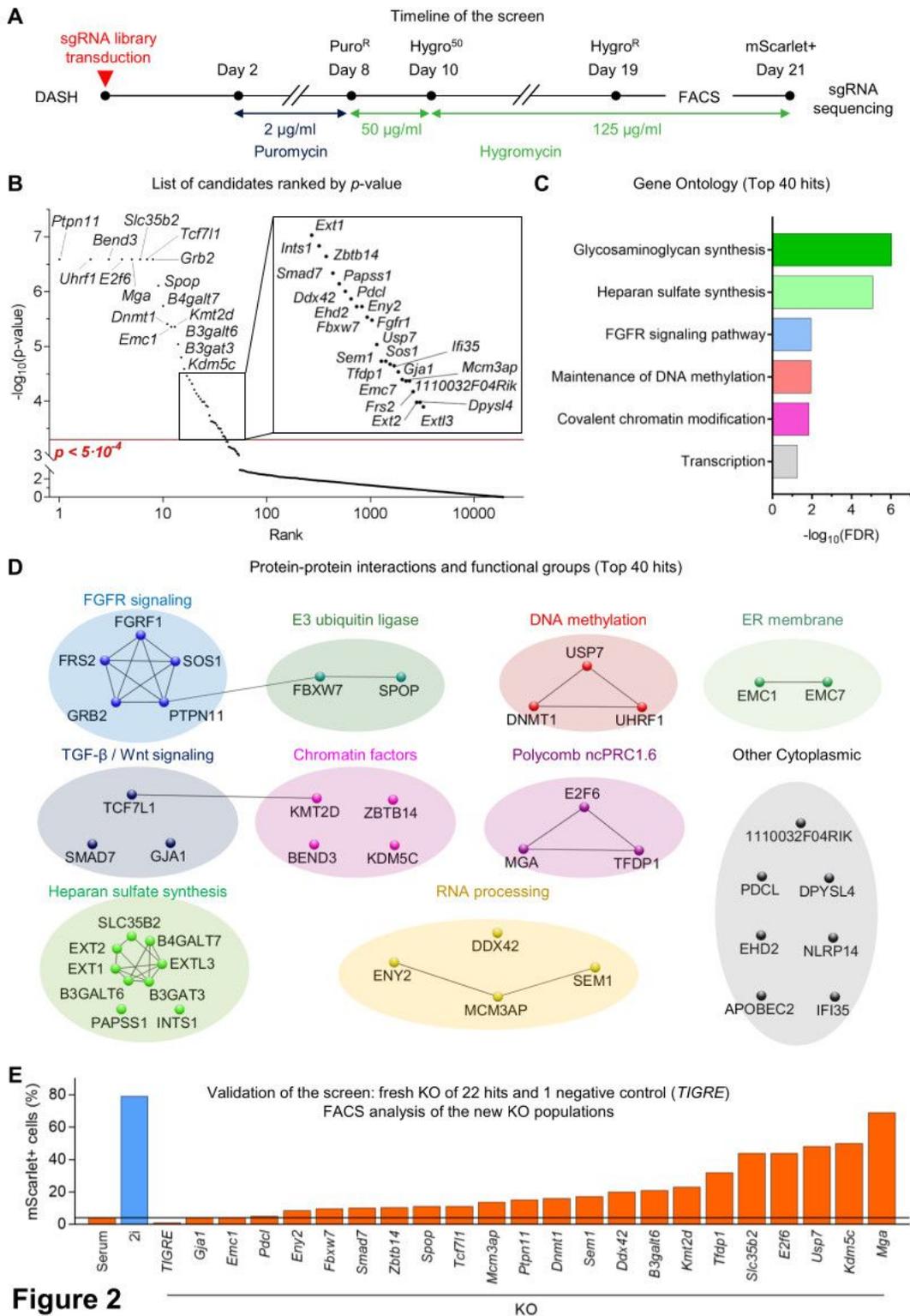
## Figures



**Figure 1**

## Figure 1

***Dazl* as a reporter gene for the 2-cell stage, screen design and generation of the *Dazl*-mScarlet-Hygro<sup>R</sup> (DASH) reporter cell line.** (A) *Dazl* is highly expressed in 2-cell-like cells *in vitro* (analysis of RNA-seq data from the indicated references). (B) *Dazl* expression in 2i is similar to the late 2-cell embryo, while expression in Serum is similar to the 8-cell stage (embryo data from GSE66390<sup>56</sup>, ESC data: this study) (C) Principle of the screen. Knocking out factors essential for repression should trigger gene reactivation in Serum. Unmethylated CpG: white lollipop; methylated CpG: black lollipop. Permissive chromatin: white circle; repressive chromatin: gray circle. (D) The DASH reporter ES cell line. A reporter cassette is inserted into one of the *Dazl* alleles, it encodes two proteins separated by 2A self-cleaving peptides (P2A, T2A): the red fluorescent mScarlet and the Hygromycin resistance enzyme (Hygro<sup>R</sup>). (E) DASH cells express mScarlet in 2i, but not in Serum. FACS analysis (50,000 cells per condition in this panel and all following panels, n=3 biological replicates). (F) DASH cells are resistant to Hygromycin in 2i, but not in serum. Surviving cells are stained with crystal violet.

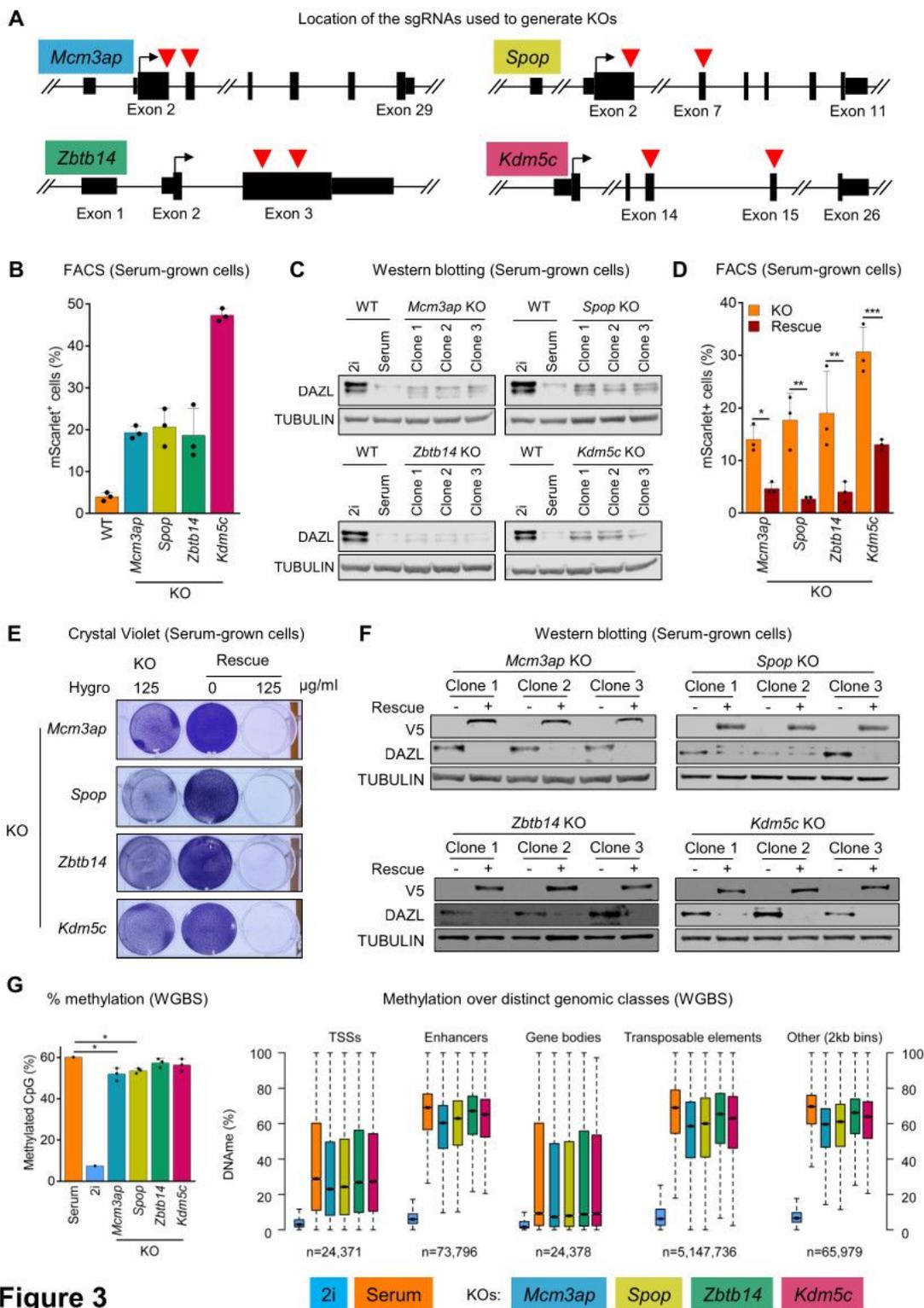


**Figure 2**

**Figure 2**

**A genome-wide CRISPR knock-out screen for *Dazl* gene reactivation in the DASH reporter line.** (A) Timeline of the screen. (B) Ranking of the hits by  $p$ -value. The top 40 hits ( $p < 0.0005$ ) are above the red line. (C) Gene ontology (GO) terms (Biological Process) significantly enriched among the top 40 hits. (D) Protein-protein interaction network clustering of the top 40 hits, generated by STRING analysis. (E) FACS analysis

of mScarlet expression after KO of the indicated candidate genes. The parental DASH cells (cultured in Serum or 2i), and cells expressing a sgRNA targeting the inert *TIGRE* locus served as controls.

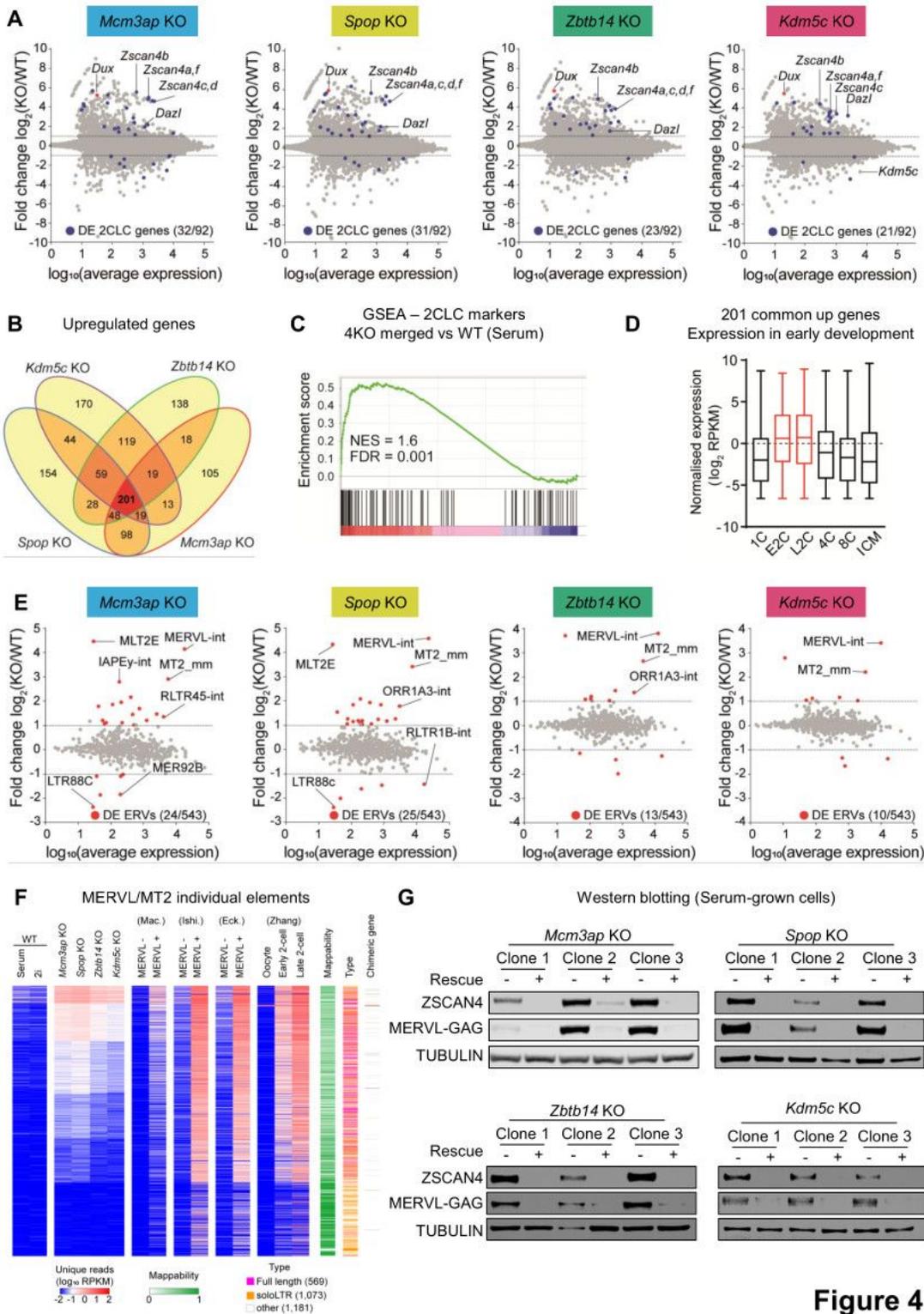


**Figure 3**

**Figure 3**

**Four loss-of-function mutants that induce Dazl expression: Mcm3ap KO, Spop KO, Zbtb14 KO, and Kdm5c KO.** (A) Depiction of the gene structures and position of the sgRNAs used to generate mutations (red

arrowheads). (B) FACS analysis of mScarlet expression after KO of the indicated genes (3 independent KO clones for each gene). (C) Western blot analysis of *Dazl* expression in the indicated cells. (D) FACS analysis: genetic rescue of each KO suppresses mScarlet expression. \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$  (Sidak post-hoc test following ANOVA). (E) Genetic rescue of each KO suppresses Hygromycin resistance. (F) Western blot analysis: genetic rescue of each KO suppresses DAZL protein induction. Each rescue construct contained a V5 tag. (G) Whole-genome bisulfite sequencing (WGBS) detects varying decreases of DNA methylation in the different KO populations. Left panel: global levels of CpG methylation. \* $p < 0.05$  (two-tailed  $t$ -test). Right panel: analysis at the indicated genomic regions. In the boxplots, the thick line indicates the median, the box limits indicate the upper and lower quartiles, and the whiskers extend to 1.5 IQR from the quartiles. The color code is indicated below the figure.

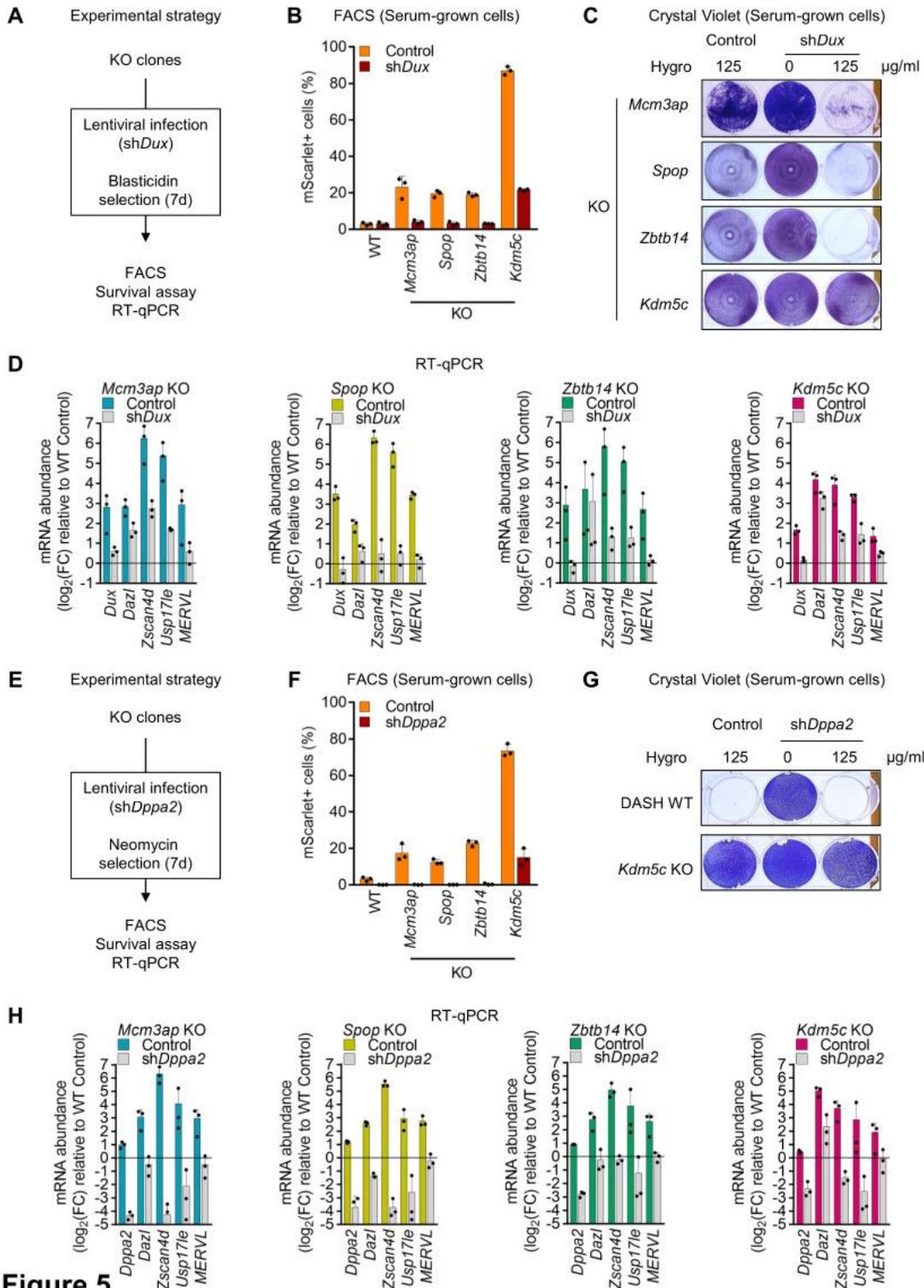


**Figure 4**

**Figure 4**

**Induction of a 2-cell-like signature in the *Mcm3ap*, *Spop*, *Zbtb14*, and *Kdm5c* KOs.** (A) RNA-seq analysis of the indicated KOs (3 independent clones each). The MA-plots are shown, with *Dux* depicted in red, and the differentially expressed (DE) genes of the 2CLC signature<sup>15</sup> depicted in blue. (B) Venn diagram of upregulated genes (RNA-seq) after knock-out of the indicated genes. (C) The 2CLC signature<sup>15</sup> is significantly enriched in the transcriptome of the 4 KOs. (D) Boxplots of normalized expression (RPKM)

for the 201 shared upregulated genes in early developmental stages. In the boxplots, the line indicates the median, the box limits indicate the upper and lower quartiles, and the whiskers extend to 1.5 IQR from the quartiles. Transcriptome data was obtained from GSE66390<sup>56</sup>. E2C: early 2-cell; L2C: late 2-cell. (F) MA-plots for repeated elements in our RNA-seq data (3 independent KO clones for each gene). (F) Heatmap depicting the normalized expression (RPKM) of MERVL/MT2 single elements (full length, solo LTR, or others) in 4 candidates KO (3 independent clones each) and parental DASH ESCs. Mappability score and the presence of chimeric genes are indicated. Published data of MERVL-positive cells and 2C embryos were included for comparison and were obtained from GSE33923<sup>7</sup> (Macfarlan 2012); E-MTAB-2684<sup>57</sup> (Ishiuchi 2015); GSE75751<sup>23</sup> (Eckersley-Maslin 2016) and GSE71434<sup>58</sup> (Zhang 2016). (G) Induction of the ZSCAN4 and MERVL-GAG proteins in the KOs, and inhibition in the rescue clones.



**Figure 5**

**Figure 5**

**Epistasis analyses show that KDM5C removal activates Dazl independently of Dux and Dppa2.** (A) Experimental scheme for *Dux* depletion. (B) FACS analysis after *shDux* shows that depleting *Dux* brings back the expression of mScarlet to WT levels in all KO, except for *Kdm5c* KO. (C) Depleting *Dux* cancels Hygromycin resistance in all KO, except for *Kdm5c* KO. (D) RT-qPCR analysis: the induction of 2CLC markers is *Dux* dependent, except for *Dazl* induction in the *Kdm5c* KO. (E) Experimental scheme for *Dppa2*

depletion. (F) FACS analysis after sh*Dppa2* shows that depleting *Dppa2* brings back the expression of mScarlet to WT levels in all KOs, except for *Kdm5c* KO. (G) Depleting *Dppa2* cancels Hygromycin resistance in all KOs, except for *Kdm5c* KO. (H) RT-qPCR analysis: the induction of 2CLC markers is *Dppa2* dependent, except for *Dazl* induction in the *Kdm5c* KO.

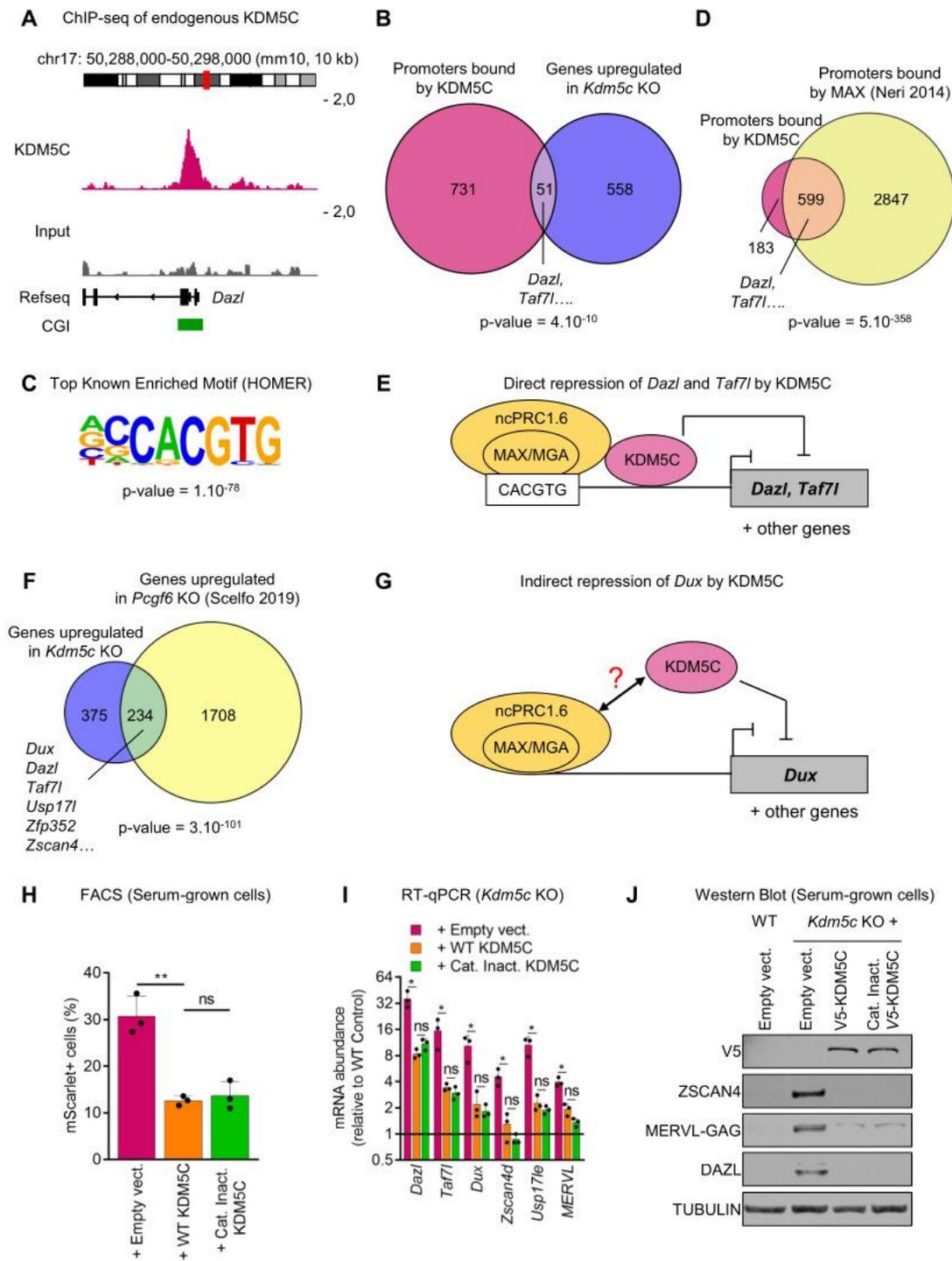
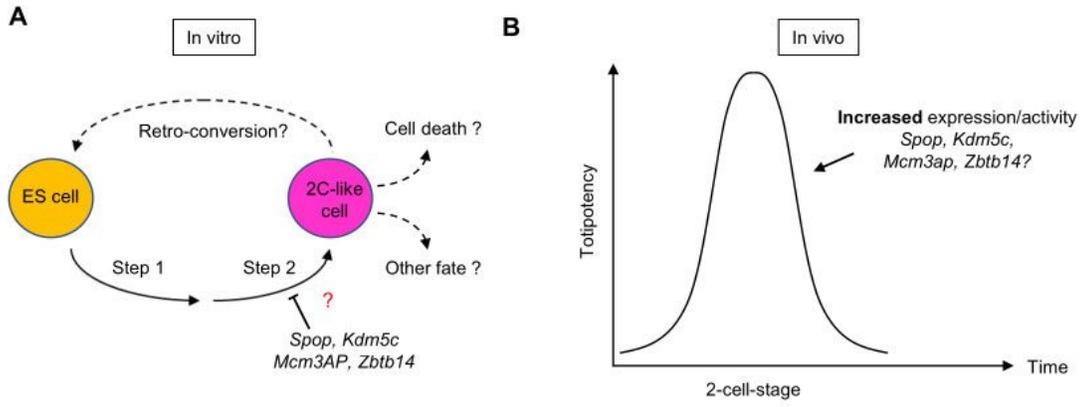


Figure 6

Figure 6

**Catalysis-independent roles of KDM5C.** (A) KDM5C binds the *Dazl* promoter. ChIP-seq of endogenous KDM5C in DASH cells. (B) KDM5C is necessary for the repression of some of its promoter targets, including *Dazl*. p-value: hypergeometric test (for this panel and all following diagrams). (C) The top known binding motif called by HOMER for KDM5C peaks is the E-box, which is the target of the MAX/MGA dimer. (D) 77% (599/782) of all promoters bound by KDM5C are also bound by MAX, a DNA-binding subunit of ncPRC1.6 (data from GSE48175). (E) Among the 2CLC genes, *Dazl* and *Taf7l* are directly bound and repressed by KDM5C and ncPRC1.6. (F) 38% (234/609) of genes upregulated in *Kdm5c* KO are also upregulated in *Pcgf6* KO, including 2CLC markers (data from GSE122715). (G) Representation of the role of KDM5C at *Dux*. A possible mechanism for KDM5C-mediated repression at *Dux* would be the recruitment by ncPRC1.6. (H) The WT version of KDM5C efficiently represses mScarlet expression in the *Kdm5c* KO, and so does the catalytically inactive (H514A) mutant. (I) RT-qPCR analysis: the WT version of KDM5C efficiently represses 2CLC gene expression in the *Kdm5c* KO, and so does the catalytically inactive mutant. (J) Western-blot analysis: the WT version of KDM5C efficiently represses 2CLC protein expression in the *Kdm5c* KO, and so does the catalytically inactive mutant.



**Figure 7**

Figure 7

*Summary and hypotheses.*

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- [FigureS1.jpg](#)
- [FigureS2.jpg](#)
- [FigureS3.jpg](#)
- [FigureS4.jpg](#)
- [FigureS5.jpg](#)
- [FigureS6.jpg](#)
- [SupplementaryFile1Oligonucleotidesequences.xlsx](#)
- [SupplementaryFile2Screenresults.xlsx](#)
- [SupplementaryFile3Datasetslist.xlsx](#)
- [SupplementaryFile4WGBSbasicmetrics.xlsx](#)
- [Supplementaryfile5RNAseqdifferentiallyexpressedgenes.xlsx](#)
- [Supplementaryfile6RNAseqGSEA.xlsx](#)